McGRAW-HILL

SERIES IN 225

PROBABILITY

AND STATISTICS

STATISTICAL DECISION THEORY

McGraw-Hill Series in Probability and Statistics

DAVID BLACKWELL, *Consulting Editor*

BHARUCHA-REID. Elements of the Theory of Markov Processes and Their Applications

GRAYBILL. Introduction to Linear Statistical Models, Volume I

WADSWORTH AND BRYAN. Introduction to Probability and Random Variables

WEISS. Statistical Decision Theory

STATISTICAL DECISION THEORY

LIONEL WEISS

ASSOCIATE PROFESSOR
DEPARTMENT OF INDUSTRIAL AND
ENGINEERING ADMINISTRATION
CORNELL UNIVERSITY

New York Toronto London 1961
McGRAW-HILL BOOK COMPANY, INC.

STATISTICAL DECISION THEORY

Library of Congress Catalog Card Number 61-11132

69068

PREFACE

At the present time there are very few textbooks on statistical decision theory, and there is apparently no textbook that gives a reasonably complete discussion of decision theory at an intermediate mathematical level. The author hopes that this textbook will help to fill the gap. It has as formal mathematical prerequisites only calculus through partial differentiation and multiple integration, plus some elementary facts about the use of determinants in solving linear equations.

A few sections of the text are demanding on a student with the amount of mathematical maturity usually implied by a year course in calculus. The derivation of the Wald sequential decision rule in Chapter 7 is the most important example. The student will be familiar with all the mathematical tools used, but because of the length of the development, the instructor will have to keep the student from being overwhelmed by the details.

Although the purpose of the text is to teach statistical decision theory, a substantial proportion of the text is devoted to a simple and relatively brief discussion of probability theory. This makes the text self-contained for those students who have not previously studied probability theory. Students who have had a course in probability theory should find the discussion of this theory a useful review, since the emphasis is on those topics of probability theory most useful in statistical decision theory.

Some of the topics discussed in the text are unusual for courses in the general field of statistics. Two examples are linear programming and making a sequence of nonsampling decisions over time. The author feels that these topics belong in a course on statistical methods.

For reasons explained in Chapter 5, the loss is made to depend on the decision chosen and on chance variables that will be observed after the decision has been chosen, rather than on the decision chosen and the distribution of the chance variables.

Chapter 9, the last chapter, introduces conventional statistical methods as special cases of statistical decision theory. The examples discussed in Chapter 9 seem a step further from practical problems than the examples discussed in earlier chapters. This seems to be an inherent property of the problems discussed in conventional statistical theory.

In a text at this level, it does not seem desirable to include detailed statements about which researchers are responsible for the various results described. Statistical decision theory is the creation of the late A. Wald; the proof that the Wald sequential rule is a Bayes decision rule is due to A. Wald and J. Wolfowitz; and the simplex method is due to G. B. Dantzig.

How long a course should be devoted to the text depends on the maturity of the students and on the taste and style of the instructor. The author has used forty-five hours of lectures for a brief review of Chapters 1 to 4 and a thorough discussion of the remaining chapters, including the discussion of many numerical examples. This will seem to be a slow pace to many instructors, who will be able to cover the whole text fairly thoroughly in a forty-five-hour course.

Chapter 9 serves as an introduction to conventional statistical theory, and it is possible to use the text for one semester and then to use a text on conventional statistical methods for the second semester of a year course.

The author is indebted to Professor Sir Ronald A. Fisher, F.R.S., Cambridge, and to Messrs. Oliver & Boyd Ltd., Edinburgh, for permission to reprint the tables of the chi-square and the *t* distribution from "Statistical Methods for Research Workers." The author is also indebted to Professor G. W. Snedecor and the Iowa State University Press, Ames, Iowa, for permission to reprint the table of the *F* distribution from "Statistical Methods."

For valuable typing assistance, the author is grateful to Miss Ruth Ritchie of the University of Virginia, Mrs. David Freeman, formerly of Ithaca, N.Y., Mrs. Robert Berlin of Syracuse, N.Y., and to Rhoda Weiss.

Lionel Weiss.

CONTENTS

Chapter 1

ELEMENTARY PROBABILITY THEORY

1.1. Introduction. Everything in this text is based on the mathematical theory of probability; therefore a simple but detailed development of this theory will be given. The theory will be developed in two separate ways. The first way is an intuitive approach; the second way is an axiomatic approach. The axiomatic method is neater and mathematically more satisfactory. However, fruitful applications of probability theory to the physical world are based on our intuitive understanding of both probability theory and the physical world. Therefore it is essential to develop the proper intuitive feeling for probability theory.

1.2. Intuitive Definition of Probability. When we say "the probability of getting a head when tossing a perfectly balanced coin is equal to ½," we mean that when such a coin is tossed a large number of times, a head will come up on approximately one-half of the tosses.

In general, the statement "the probability of getting the event E in one trial is equal to p" means that in a large number of similar trials, the proportion of trials in which E will occur will be close to p.

Note that in the definition we did not say how large the number of trials should be, nor how close the proportion of occurrences of E will be to p. All that we can say about this now is that intuitively we feel that the larger the number of trials, the closer the proportion of occurrences will be to the probability.

There is another aspect of our definition. If a coin when tossed gives an exactly alternating sequence of heads and tails, extending indefinitely as follows: head, tail, head, tail, head, tail, . . . , it is true that the proportion of times a head appears approaches ½ as the number of tosses increases, but there is too much regularity here for us to apply our simple probability theory. In other words, the trials on which a head occurs must be scattered irregularly among all the trials. Another way of putting this is to say that we should be able to forecast the long-run proportion of heads *only* and nothing else.

In practice, we usually do not know the numerical value of the probability of a given event. Experimentation has shown that the probability of getting a head when tossing an ordinary coin is *not* exactly ½. Presumably, the different designs on the two sides of the coin throw it slightly off perfect balance. By definition, a "true" or "fair" or "well-balanced" coin is one with the probability of a head equal to ½, but such a coin may not actually exist in the physical world. Similarly, a "fair" die is one with the probability of each face equal to ⅙. Such a die may not exist in the physical world, but that need not prevent us from developing a satisfactory theory based on the assumed existence of such a fair die.

1.3. Use of the Intuitive Definition of Probability to Develop the Basic Rules of Probability Theory. The identification of probabilities as approximately equal to proportions of occurrences will enable us to develop certain basic rules of probability theory.

The symbol $P(E)$ is to be read "the probability of the event E."

Since, for any event, the proportion of times it occurs in any sequence of trials is between 0 and 1, it is clear that any probability must be contained within the limits 0 and 1.

If E is any event, we define the event "not E" as that event which occurs on each trial where E fails to occur. Thus, if E is the appearance of a head on a toss of a coin, then (not E) is the appearance of a tail, assuming the coin cannot stand on its edge. Suppose we observe N trials and let $n(E)$ denote the number of trials on which E occurs, while $n(\text{not } E)$ denotes the number of trials on which (not E) occurs. We must have $n(E) + n(\text{not } E) = N$. Dividing through by N, we get $n(E)/N + n(\text{not } E)/N = 1$. But by our definition of probability, $n(E)/N$ is close to $P(E)$, $n(\text{not } E)/N$ is close to $P(\text{not } E)$. Thus we have as a basic rule of probability: for any event E, $P(E) + P(\text{not } E) = 1$.

If D, E are any events, we define the event "D and E" as that event which occurs on each trial where both the events D and E occur. Thus, if D is the appearance of a spade when turning up the top card in a deck while E is the appearance of an ace when turning up the top card, the event (D and E) is the appearance of the ace of spades when turning up the top card.

If D, E are any events, we define the event "D or E" as that event which occurs on each trial where at least one of the events D, E occurs. Thus, if D is the appearance of a spade when turning up the top card in a deck while E is the appearance of an ace, the event (D or E) is the appearance of an ace or a spade (any of 16 separate cards) as the top card.

Suppose D, E are any two events and we observe N trials. $n(\ )$ denotes the number of trials on which the event in parentheses occurs. Then we have $n(D \text{ or } E) = n(D) + n(E) - n(D \text{ and } E)$. To see this, note that $n(D) + n(E)$ would be greater than $n(D \text{ or } E)$ by exactly $n(D \text{ and } E)$, since those trials on which both D and E occurred would be counted in both $n(D)$ and $n(E)$. The following numerical example illustrates this. Suppose each trial consists of shuffling a full deck of cards and turning up the top card. The event D is turning up either a jack or a queen, while the event E is turning up either a queen or a king. Then the event (D or E) is turning up a jack, queen, or king, while the event (D and E) is turning up a queen. Suppose 1,000 trials are observed and a jack appears on 72 trials, a queen on 76 trials, and a king on 77 trials. Then $n(D) = 72 + 76$, $n(E) = 76 + 77$, $n(D \text{ and } E) = 76$, $n(D \text{ or } E) = 72 + 76 + 77$, illustrating the fact that $n(D \text{ or } E) = n(D) + n(E) - n(D \text{ and } E)$. Dividing this last equality through by N, we get

$$\frac{n(D \text{ or } E)}{N} = \frac{n(D)}{N} + \frac{n(E)}{N} - \frac{n(D \text{ and } E)}{N}$$

But in this equality, each ratio is the proportion of times an event occurs, and these proportions are close to the corresponding probabilities, by our definition of probability. This gives another basic rule of probability theory: for any events $D, E, P(D \text{ or } E) = P(D) + P(E) - P(D \text{ and } E)$.

Whenever the events D, E are such that it is impossible for them both to occur on the same trial, they are said to be "mutually exclusive." As an example, a trial might consist of turning up the top card of a deck, D might be the appearance of a heart as the top card, E might be the appearance of a black card as top card. If D, E are mutually exclusive, then $P(D \text{ and } E) = 0$, and $P(D \text{ or } E) = P(D) + P(E)$.

1.4. Intuitive Definition of Conditional Probability. We have defined the probability of an event E as the approximate proportion of trials on which E occurs in a long series of trials. Now we define the "conditional probability of E given that D occurs" as the approximate proportion of trials where E occurs *among those trials where D occurs;* that is, in computing the proportion, we disregard all trials where D does not occur. The symbol $P(E \mid D)$ denotes the conditional probability of E, given that D occurs.

Suppose we observe N trials and $n(\ )$ denotes the number of trials where the event in parentheses occurs. The proportion of trials where E occurs, *among those trials where D occurs*, is equal to

$$\frac{n(E \text{ and } D)}{n(D)}$$

which is equal to

$$\frac{n(E \text{ and } D)/N}{n(D)/N}$$

But this last expression is close to $P(E \text{ and } D)/P(D)$. Thus we are led to define $P(E \mid D)$ as $P(E \text{ and } D)/P(D)$. [Here we assume $P(D) > 0$.]

1.5. Independent Events. The statement "the event E is independent of the event D" is defined to mean that $P(E \mid D) = P(E)$. The reason for the use of the word "independent" here is that the probability of E remains the same whether we know that D occurs or not; so knowledge about D does not change our knowledge about E.

Our definition of $P(E \mid D)$ implies that $P(E \text{ and } D) = P(D)P(E \mid D)$. If E is independent of D, we then have $P(E \text{ and } D) = P(E)P(D)$. Conversely, if $P(E \text{ and } D) = P(E)P(D)$, we have

$$P(E \mid D) = \frac{P(E \text{ and } D)}{P(D)} = \frac{P(E)P(D)}{P(D)} = P(E)$$

so E is independent of D. Thus we could define E's independence of D by the equality $P(E \text{ and } D) = P(E)P(D)$.

Similarly, we define "D is independent of E" to mean $P(D \mid E) = P(D)$, and we find that this is equivalent to $P(E \text{ and } D) = P(E)P(D)$. This means that if D is independent of E, then E is independent of D, and vice versa, so it is customary to say simply that "D and E are independent," which is equivalent to any of the equalities $P(D \mid E) = P(D)$, $P(E \mid D) = P(E)$, $P(D \text{ and } E) = P(D)P(E)$.

As an example of independent events, suppose a trial consists of thoroughly shuffling a full deck of cards and turning up the top card. D is the appearance of a spade as top card; E is the appearance of a picture card as top card. By the intuitive meaning of "thorough shuffling," we have $P(D) = 13/52$, $P(E) = 12/52$, and $P(D \text{ and } E) = 3/52$. Since in this case $P(D \text{ and } E) = P(D)P(E)$, the events D and E are independent.

As an example of events which are not independent, suppose a trial consists of rolling a fair die once; the event D is the appearance of an even number of spots, while the event E is the appearance of fewer than four spots. Then $P(D) = 3/6$, $P(E) = 3/6$, $P(D \text{ and } E) = 1/6$, and since $P(D \text{ and } E)$ is not equal to $P(D)P(E)$, the events D and E are not independent in this case.

When the occurrence of the event D depends on an experiment that has no physical connection with the experiment determining the occurrence of the event E, it is always assumed that the events D and E are independent. Thus, if we toss a penny and a nickel and D is the appearance of a head on the penny and E is the appearance of a tail on the nickel, D and E are independent events.

1.6. Additional Basic Rules of Probability. Given any r events $E_1, \ldots, E_r$, the event "E_1 or E_2 or $\cdots$ or E_r" is defined as that event which occurs on any trial where at least one of the r events occurs. The event "E_1 and E_2 and $\cdots$ and E_r" is defined as that event which occurs on any trial where all r events occur.

The events $E_1, E_2, \ldots, E_r$ are called "mutually exclusive by pairs" if no two of them can occur on the same trial. If N trials are made, and if $E_1, \ldots, E_r$ are mutually exclusive by pairs, we must have $n(E_1 \text{ or } E_2 \text{ or} \cdots \text{or } E_r) = n(E_1) + n(E_2) + \cdots + n(E_r)$, and dividing through by N, we get

$$\frac{n(E_1 \text{ or } E_2 \text{ or} \cdots \text{ or } E_r)}{N} = \frac{n(E_1)}{N} + \frac{n(E_2)}{N} + \cdots + \frac{n(E_r)}{N}$$

Since each ratio is close to the corresponding probability, we get the following basic rule: if $E_1, E_2, \ldots, E_r$ are any events which are mutually exclusive by pairs, then $P(E_1 \text{ or } E_2 \text{ or} \cdots \text{or } E_r) = P(E_1) + P(E_2) + \cdots + P(E_r)$.

If $D_1, D_2, \ldots, D_r$ are any events, they are called "mutually independent" if $P(D_a \text{ and } D_b \text{ and } \cdots \text{ and } D_n) = P(D_a)P(D_b) \cdots P(D_n)$, where $a, b, \ldots, n$ are any integers all different from each other and all between 1 and r. This definition of independence is a generalization of the definition of the independence of two events. If each of the events $D_1, D_2, \ldots, D_r$ is defined by an experiment physically separated from the experiments determining the other D's, then $D_1, D_2, \ldots, D_r$ are assumed to be mutually independent.

1.7. Axiomatic Development of Elementary Probability Theory. All the rules of probability that we derived above can be derived very easily from a simple axiom system. The following is a sketch of this approach.

We start with a given finite number k of "fundamental occurrences," which we represent by the symbols $F_1, F_2, \ldots, F_k$. The mathematical theory says no more to define these fundamental occurrences, but they are meant to correspond to all possible mutually exclusive outcomes of an experiment. Thus, if the experiment is rolling a die, $k = 6$ and each fundamental occurrence corresponds to one of the six faces of the die. If the experiment is picking the top card from a deck of cards, $k = 52$ and each fundamental occurrence corresponds to one of the 52 cards. If the experiment is picking the top five cards from a deck of cards, $k = (52)(51)(50)(49)(48)$ and each fundamental occurrence corresponds to one of the ways of picking a sequence of five different cards.

Attached to each fundamental occurrence is a nonnegative number. The number attached to F_i will be denoted by p_i and is called "the probability of F_i." $p_1 + p_2 + \cdots + p_k = 1$. These p's may be assigned arbitrarily, provided they are nonnegative and sum to unity.

However, p_i is of course meant to correspond to the proportion of times F_i will occur in a long sequence of performances of the experiment, and in any application of the theory p_i is set accordingly. Thus, if we apply the theory to the experiment which consists of rolling a well-balanced die, $k = 6$ and each p_i is set equal to $1/6$. In fact, this would be the mathematical definition of "well-balanced die."

Any given set of fundamental occurrences is called an "event." The probability of any event is defined as the sum of the probabilities attached to all the fundamental occurrences in the event. (In listing the fundamental occurrences in an event, it is important to list each only once, to avoid double counting.) An event is considered to occur on any trial where one of the fundamental occurrences in the event occurs, and this is consistent with the definition of the probability of the event.

The set consisting of *no* fundamental occurrences is also an event (often called the "impossible event"), and its probability is defined as zero.

Given any event C, the event (not C) is defined as the set of all fundamental occurrences which are not in the event C. From this, it follows immediately that $P(C) + P(\text{not } C) = 1$.

If D, E are any two events, the event (D or E) is defined as the set of all fundamental occurrences appearing in either of the events D or E or in both, while the event (D and E) is defined as the set of all fundamental occurrences appearing in both the event D and the event E. From these definitions, it is easily seen that the event (D or E) occurs whenever at least one of the events D, E occurs, while the event (D and E) occurs whenever both events D, E occur.

Two events A, B are called "mutually exclusive" if there is no fundamental occurrence which is in both A and B. This means that the event (A and B) contains no fundamental occurrences, so $P(A \text{ and } B) = 0$. Similarly, the events $A_1, A_2, \ldots, A_n$ are called "mutually exclusive by pairs" if there is no fundamental occurrence which is in more than one of the events $A_1, A_2, \ldots, A_n$.

Suppose we represent each fundamental occurrence as a cross in a diagram and represent an event by enclosing the set of fundamental occurrences comprising the event (Fig. 1.1). The event (D or E) consists of all the fundamental occurrences in the hatched portion, while the event (D and E) consists of all the fundamental occurrences in the double-hatched portion. Remembering that the probability of an event is the sum of the probabilities attached to the fundamental occurrences in the event (counting each fundamental occurrence only once), it is easily seen from the diagram that $P(D \text{ or } E) = P(D) + P(E) - P(D \text{ and } E)$.

Defining the event (B or C or D) as the set of fundamental occurrences appearing in at least one of the events B, C, D and defining the event (B and C and D) as the set of fundamental occurrences appearing in all three events B, C, D, a diagram similar to Fig. 1.1 can be used to prove the following formula:

$$P(B \text{ or } C \text{ or } D) = P(B) + P(C) + P(D) - P(B \text{ and } C)$$
$$- P(B \text{ and } D) - P(C \text{ and } D) + P(B \text{ and } C \text{ and } D)$$

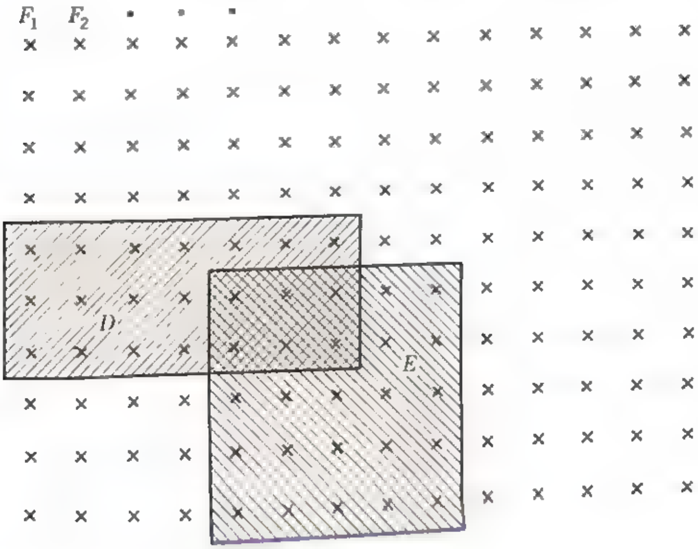


Fig. 1.1

1.8. Examples of the Applications of the Basic Rules of Probability. In this section we discuss three examples of probability calculations involving a fairly elaborate use of the basic rules of probability theory.

Example 1. Suppose we have a coin with probability p of coming up head on each single toss, and we toss the coin m times. We wish to find the probability that a head will appear on exactly k of the tosses, where k is a given integer between 0 and m.

Here a trial consists of m separate tosses. We keep track of what happens on a trial by listing the result of the first toss, the second toss, . . . , the mth toss. Thus there are 2^m different possible outcomes of the experiment, each outcome being described by a sequence of H's (for heads) and T's (for tails); m symbols in all. Thus the sequence

$$\overbrace{HH\cdots HH}^{m}$$

represents the appearance of a head on every one of the m tosses; the sequence

$$\overbrace{H\cdots H}^{k}\quad\overbrace{T\cdots T}^{m-k}$$

represents the appearance of a head on each of the first k tosses, and a tail on each of the last $m - k$ tosses. Each of these sequences represents a fundamental occurrence.

Our next task is to assign a probability to each fundamental occurrence. This should be done in a way that takes account of the description of the physical circumstances of the problem. Since there is no physical connection between the different tosses of the coin, an event defined in terms of the first toss is independent of an event defined in terms of the second toss, or the third toss, etc. Then, by the rule of Sec. 1.6, we should find the probability of the fundamental occurrence $HH \cdots HH$, for example, by noting that this fundamental occurrence is the occurrence of (head on toss 1 and head on toss 2 and $\cdots$ and head on toss m), and the probability of this is P(head on toss 1)P(head on toss 2) $\cdots$ P(head on toss m). But P(head on toss i) was specified as p, and therefore the probability of the fundamental occurrence $HH \cdots HH$ is set at p^m. By the same sort of reasoning, any fundamental occurrence with exactly r H's (and therefore $m - r$ T's) is assigned the probability $p^r(1 - p)^{m-r}$ (recalling that the probability of a tail on any given toss is $1 - p$). This gives the complete assignment of probabilities to fundamental occurrences.

Now we return to our original problem, finding the probability of getting exactly k heads. The event (k heads occur) consists of all fundamental occurrences containing exactly k H's, and P(k heads occur) is equal to the sum of the probabilities attached to these fundamental occurrences. There are $m!/[k!\,(m - k)!\,]$ such fundamental occurrences, one for each way of choosing k places out of m places, and each such fundamental occurrence has been assigned the probability $p^k(1 - p)^{m-k}$. Thus we have

$$P(k \text{ heads occur}) = \frac{m!}{k!\,(m-k)!}\, p^k(1-p)^{m-k}$$

Example 2. We choose the top five cards from a well-shuffled deck of cards. We want to find the probability that these five cards consist of three spades and two clubs.

We keep track of what happens on a trial by listing the top card, the second card, third card, fourth card, fifth card, in that order. Thus there are (52) (51) (50) (49) (48) different possible outcomes of the experiment, each being a fundamental occurrence.

The natural way to assign probabilities to fundamental occurrences in this case is to assign the same probability to each fundamental occurrence. This is so because there is no reason to expect one particular outcome to occur more or less often than some other particular outcome: this is the intuitive meaning of "well-shuffled deck." Thus we assign the probability $[(52)\,(51)\,(50)\,(49)\,(48)]^{-1}$ to each fundamental occurrence, to make the sum of the probabilities equal to 1.

To find the desired probability of getting three spades and two clubs, it is merely necessary to multiply the number of fundamental occurrences in this event by the common probability assigned to each fundamental occurrence. We find the number of fundamental occurrences in the event as follows. For each specification of three particular spades and two particular clubs, there are 5! different fundamental occurrences containing those particular cards, since each permutation of the cards is a different fundamental occurrence. But there are 13!/(3! 10!) different ways of choosing 3 spades out of the 13 spades and 13!/(2! 11!) of choosing 2 clubs out of the 13 clubs. Therefore the number of different fundamental occurrences in the event is

$$5!\,\frac{13!}{3!\,10!}\,\frac{13!}{2!\,11!}$$

Example 3. There are two boxes, labeled I, II. Box I contains 3 red cards and 4 black cards; box II contains 7 red cards and 5 black cards. The cards in box I are shuffled thoroughly; then a card is drawn from box I and (without being observed) is transferred to box II. Then the cards in box II are shuffled thoroughly, and one is drawn. The problem is to find the probability that the card drawn from box II is red.

In order to keep track of the fundamental occurrences in this example, we imagine that each card is labeled as follows. The 3 red cards in box I are labeled Ir1, Ir2, Ir3, respectively; the 4 black cards in box I are labeled Ib1, Ib2, Ib3, Ib4, respectively; the 7 red cards in box II are labeled IIr1, IIr2, . . . , IIr7, respectively; the 5 black cards in box II are labeled IIb1, IIb2, . . . , IIb5, respectively. Then a fundamental occurrence is described by listing a card whose label starts with I, and then listing either the same card or a card whose label starts with II: the first card listed is the card transferred from box I to box II; the second card listed is the card finally drawn from box II. Thus there are (7)(13) = 91 different fundamental occurrences.

Our next task is to assign a probability to each fundamental occurrence, in a way that conforms to the physical description of the experiment. Let us examine the fundamental occurrence (Ir1,Ir1). This means that Ir1 is transferred, and then Ir1 is drawn. The probability that Ir1 is transferred is equal to $\frac{1}{7}$, by the intuitive meaning of thorough shuffling. Again by the intuitive meaning of thorough shuffling, the conditional probability that Ir1 is drawn from box II, given that Ir1 is transferred from box I, is $\frac{1}{13}$. From Sec. 1.4, $P(D \text{ and } E) = P(D)P(E \mid D)$, for any events D, E. Denoting by D the event (Ir1 is transferred from box I) and by E the event (Ir1 is drawn from box II), we find that $P(D \text{ and } E) = P(\text{Ir1,Ir1}) = (1/7)(1/13) = 1/91$. Thus the natural probability to assign to the fundamental occurrence (Ir1,Ir1) is $1/91$. Exactly the same reasoning shows that the natural probability to assign to each of the 91 fundamental occurrences is $1/91$.

Next we count the number of different fundamental occurrences in the event (card drawn from box II is red). There are 8 different fundamental occurrences in this event starting with Ir1, 8 starting with Ir2, 8 starting with Ir3, 7 starting with Ib1, 7 starting with Ib2, 7 starting with Ib3, 7 starting with Ib4. Altogether, this is a total of $(3)(8) + (4)(7) = 52$ different fundamental occurrences in the event. Therefore P(card drawn from box II is red) $= 52/91$.

It is interesting to note that this probability can be found by the following short intuitive argument. When we draw from box II, there are 13 cards in the box, and on the average $7\frac{3}{7}$ of these cards are red: 7 red cards originally in box II plus $\frac{3}{7}$ of a red card transferred from box I "on the average." Therefore the proportion of red cards in box II will be $7\frac{3}{7}/13$ "on the average," and since the shuffling is thorough, the probability of drawing a red card from box II is $7\frac{3}{7}/13 = 52/91$.

In many problems, counting the number of fundamental occurrences in a given event may be a very difficult task, and various ingenious formulas and methods have been developed to aid in the counting. However, our interest is in the concepts of probability theory rather than in computational techniques.

Chapter 2

CHANCE VARIABLES WITH A FINITE NUMBER OF POSSIBLE VALUES

2.1. Introduction. The fundamental occurrences of a given experiment may be any of a great variety of different objects: playing cards, faces of a coin, colors of a chip, etc. However, in most of the remainder of this textbook we shall be discussing experiments each of whose outcomes is a number, so that each fundamental occurrence is a number. Such an experiment is said to define a "chance variable."

The following intuitive definition of chance variable may be useful in developing the proper feeling for this concept: A chance variable is the number that will appear when the experiment is performed. Thus a chance variable is to be regarded as a number that has not yet been observed and is still to be chosen by a chance mechanism. A number that has already been observed is called an "observation," not a chance variable.

The terms "random variable," "stochastic variable," and "variate" are all synonyms for chance variable.

2.2. Probability Distributions. Since a chance variable is defined as a number to be determined by an experiment and is always in the future, all we can know about a chance variable is a table listing the possible values it can have, along with the respective probabilities of these values. Such a table is called the "probability distribution" of the chance variable. For example, the chance variable defined as the number that will appear when a well-balanced die is rolled has the following probability distribution:

Possible values	1	2	3	4	5	6
Probabilities	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

As another example, if our experiment is throwing a fair coin twice and the chance variable is defined as the number of throws on which a head will come up, then this chance variable has the following probability distribution:

Possible values	0	1	2
Probabilities	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

(These probabilities are derived from the first example of Sec. 1.8.)

As a matter of notation, capital letters near the end of the alphabet, such as W, X, Y, or Z, shall denote chance variables. For purposes of writing general formulas, it will be convenient to use the following notation to represent a general probability distribution for a chance variable X:

Possible values	x_1	x_2	$\cdots$	x_k
Probabilities	p_1	p_2	$\cdots$	p_k

where of course $p_1, p_2, \ldots, p_k$ are nonnegative numbers adding to 1.

Before concluding this section, we emphasize that the probability distribution of a chance variable X contains all the information it is possible to have about the chance variable. This is so because a chance variable is defined as a value to be determined by an experiment, and therefore all we can know about the chance variable are its possible values with their probabilities. From our point of view, a chance variable is described by its probability distribution rather than by a physical experiment. Rolling a well-balanced die is one physical experiment, and picking the top card from a well-shuffled deck of six cards labeled from 1 to 6 is another, but these experiments determine the same chance variable, the one whose probability distribution is

Possible values	1	2	3	4	5	6
Probabilities	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

As we shall see, statistical problems are those problems which arise when we are dealing with chance variables whose probability distributions are not completely known.

2.3. Expected Values. Suppose the chance variable X has the probability distribution

Possible values	x_1	x_2	$\cdots$	x_k
Probabilities	p_1	p_2	$\cdots$	p_k

and suppose $g(x)$ is a given function of x. Then the symbol $E\{g(X)\}$ is read "the expected value of $g(X)$" and is defined as the value $p_1g(x_1) + p_2g(x_2) + \cdots + p_kg(x_k)$.

As a numerical example, if X has the probability distribution

Possible values	-1	1
Probabilities	$\frac{1}{2}$	$\frac{1}{2}$

then $E\{X\} = \frac{1}{2}(-1) + \frac{1}{2}(1) = 0$; $E\{X^2\} = \frac{1}{2}(-1)^2 + \frac{1}{2}(1)^2 = 1$; $E\{3X + 2\} = \frac{1}{2}[3(-1) + 2)] + \frac{1}{2}[3(1) + 2] = 2$.

As another numerical example, if the probability distribution is

Possible values	1	2	3	4	5	6
Probabilities	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\begin{aligned} E\{X^2 - 2\} &= \tfrac{1}{6}(1^2 - 2) + \tfrac{1}{6}(2^2 - 2) + \tfrac{1}{6}(3^2 - 2) + \tfrac{1}{6}(4^2 - 2) \\ &\quad + \tfrac{1}{6}(5^2 - 2) + \tfrac{1}{6}(6^2 - 2) = \tfrac{79}{6} \\ E\{(X - 2)^2\} &= \tfrac{1}{6}(1 - 2)^2 + \tfrac{1}{6}(2 - 2)^2 + \tfrac{1}{6}(3 - 2)^2 + \tfrac{1}{6}(4 - 2)^2 \\ &\quad + \tfrac{1}{6}(5 - 2)^2 + \tfrac{1}{6}(6 - 2)^2 = \tfrac{31}{6} \\ E\left\{\frac{1}{X}\right\} &= \tfrac{1}{6}(\tfrac{1}{1}) + \tfrac{1}{6}(\tfrac{1}{2}) + \tfrac{1}{6}(\tfrac{1}{3}) + \tfrac{1}{6}(\tfrac{1}{4}) + \tfrac{1}{6}(\tfrac{1}{5}) + \tfrac{1}{6}(\tfrac{1}{6}) = \tfrac{49}{120} \end{aligned}$$

$E\{g(X)\}$ has an extremely important physical interpretation, which we now discuss. First we note that whatever the probability distribution of the chance variable X is, we can construct a physical experiment which defines X. For example, if we are given the probability distribution

Possible values	$-1\frac{1}{2}$	2	7	$15\frac{1}{3}$
Probabilities	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$

we can mount a "well-balanced" arrowheaded spinner on a round dial of unit circumference and mark the circumference off into arcs of lengths $\frac{1}{8}$, $\frac{1}{8}$, $\frac{1}{2}$, $\frac{1}{4}$. These arcs are labeled $-1\frac{1}{2}$, 2, 7, $15\frac{1}{3}$, respectively. The experiment is performed by spinning the spinner, the outcome being the number labeling the arc in which the arrowhead comes to rest. The chance variable defined by this experiment has the given probability distribution.

Now, given a chance variable X with the probability distribution

Possible values	x_1	x_2	$\cdots$	x_k
Probabilities	p_1	p_2	$\cdots$	p_k

suppose we construct a physical experiment defining X and perform the experiment N times, keeping a record of the results. We denote the number that comes up on the ith performance of the experiment by t_i, so that the set of numbers $t_1, t_2, \ldots, t_N$ constitutes our record of the results of the N performances of the experiment. Some of the values $t_1, t_2, \ldots, t_N$

will be equal to x_1; some will be equal to x_2; etc. We denote by $n(x_j)$ the number of values $t_1, t_2, \ldots, t_N$ which equal x_j. We have $n(x_1) + n(x_2) + \cdots + n(x_k) = N$. The average of the N quantities $g(t_1), \ldots, g(t_N)$ is

$$\frac{g(t_1) + g(t_2) + \cdots + g(t_N)}{N}$$

and by collecting equal terms in the numerator, we can write this average as

$$\frac{n(x_1)g(x_1) + n(x_2)g(x_2) + \cdots + n(x_k)g(x_k)}{N}$$

which equals

$$\frac{n(x_1)}{N}g(x_1) + \frac{n(x_2)}{N}g(x_2) + \cdots + \frac{n(x_k)}{N}g(x_k)$$

But $n(x_j)/N$ is the proportion of the N trials in which the outcome was x_j, and by our intuitive definition of probability, we expect $n(x_j)/N$ to be close to p_j, if N is large. Therefore, if N is large, we expect the average of the N quantities $g(t_1), g(t_2), \ldots, g(t_N)$ to be close to $p_1g(x_1) + p_2g(x_2) + \cdots + p_kg(x_k)$, which is $E\{g(X)\}$. Thus $E\{g(X)\}$ is a forecast of the average of the quantities $g(t_1), g(t_2), \ldots, g(t_N)$. This interpretation of an expected value as a forecast of an observed average is extremely important for statistical theory.

2.4. Cumulative Distribution Functions. If X is a chance variable with a given probability distribution, then for any given value x, we can compute $P(X \leq x)$. $P(X \leq x)$ is a function of x, and not of the chance variable X, though the form of $P(X \leq x)$ depends on the probability distribution of X. The function $P(X \leq x)$ is called the "cumulative distribution function for X."

As a numerical example, suppose X has the probability distribution

Possible values	-1	0	2
Probabilities	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$

Then it is easy to verify that

$$\begin{aligned}
P(X \leq x) &= 0 && \text{if } x < -1 \\
P(X \leq x) &= \tfrac{1}{2} && \text{if } -1 \leq x < 0 \\
P(X \leq x) &= \tfrac{1}{2} + \tfrac{1}{6} = \tfrac{2}{3} && \text{if } 0 \leq x < 2 \\
P(X \leq x) &= \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{3} = 1 && \text{if } 2 \leq x
\end{aligned}$$

The graph of $P(X \leq x)$ is given in Fig. 2.1.

In general, if the chance variable X has the probability distribution

Possible values	x_1	x_2	$\cdots$	x_k
Probabilities	p_1	p_2	$\cdots$	p_k

then the function $P(X \leq x)$ has discontinuities at the points $x_1, x_2, \ldots, x_k$, the height of the jump at x_i being equal to p_i. Between points of discontinuity, the function $P(X \leq x)$ has zero slope. From this description, it is clear that if we know the cumulative distribution function $P(X \leq x)$, we can deduce the probability distribution of X: the possible values are the points on the x axis where $P(X \leq x)$ jumps, and the corresponding probabilities are the heights of the jumps. Thus the cumulative distribution function gives just as much information about the chance variable as the probability distribution in table form, and this is all the information it is possible to have about a chance variable. The cumulative distribution function is more generally useful than the probability distribution in table form, as we shall see.

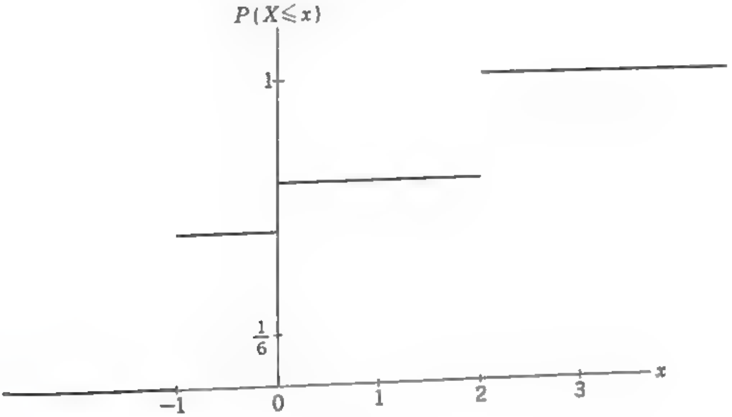


Fig. 2.1

The following formula is useful: If b, c are any values with $b < c$, then $P(b < X \leq c) = P(X \leq c) - P(X \leq b)$. To prove this, denote the event $(b < X \leq c)$ by E_1, the event $(X \leq b)$ by E_2, and the event $(X \leq c)$ by E_3. The events E_1, E_2 are mutually exclusive, and the event E_3 is exactly the same event as (E_1 or E_2). Therefore $P(E_3) = P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$, or $P(X \leq c) = P(b < X \leq c) + P(X \leq b)$, and this proves the formula.

From now on, we shall denote the phrase "cumulative distribution function" by the abbreviation "cdf." As a matter of notation, $P(X \leq x)$ will often be denoted by $F(x)$ or $G(x)$, etc.

2.5. Multivariate Chance Variables. In Sec. 2.1 we defined a chance variable as a number that will appear when an experiment is performed. Now we generalize this by considering an experiment with each possible outcome being a pair of numbers. For example, the experiment might consist of choosing a person at random from the city telephone directory and measuring his height and weight. Then each possible outcome is a pair of numbers, one giving a height, the other the weight. Or another experiment could consist of shuffling four cards and choosing the top one, each of the cards being labeled with a pair of numbers as follows: (−1,2), (3,0), (1,1), (−2,−6). In such a case, the pair of numbers that will appear when the experiment is performed is called a pair of chance variables.

Similarly, there are experiments where each possible outcome is a set of three numbers. In such a case, the set of three numbers that will appear when the experiment is performed is called a set of three chance variables. In general, there are experiments where each possible outcome is a set of *r* numbers. Then the set of *r* numbers that will appear when the experiment is performed is called a set of *r* chance variables.

2.6. Multivariate Probability Distributions. Just as in the case of a single chance variable, information about a set of chance variables is a list of the possible sets of values with their respective probabilities. For example, the pair of chance variables defined by drawing the top card in the illustration of Sec. 2.5 is completely described by the following table:

Possible pairs of values	(−1,2)	(3,0)	(1,1)	(−2,−6)
Probabilities	¼	¼	¼	¼

Suppose we denote the pair of chance variables by X, Y, with X denoting the first number of the pair that will appear when the experiment is performed, Y the second number of the pair. It will be more convenient to write the table given above in the following form:

		Possible values of X			
		−2	−1	1	3
Possible values of Y	−6	¼	0	0	0
	0	0	0	0	¼
	1	0	0	¼	0
	2	0	¼	0	0

the number in a cell being the probability that X, Y will be the pair of values in the headings for that cell. Such a table is called the "joint probability distribution of X, Y."

For purposes of writing general formulas, it will be convenient to represent the general joint probability distribution of X, Y by

		Possible values of X			
		x_1	x_2	$\cdots$	x_k
Possible values of Y	y_1	p_{11}	p_{21}	$\cdots$	p_{k1}
	y_2	p_{12}	p_{22}	$\cdots$	p_{k2}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	y_h	p_{1h}	p_{2h}	$\cdots$	p_{kh}

where $p_{ij} = P[X = x_i \text{ and } Y = y_j]$.

In the same way, we define joint probability distributions for three or more chance variables, as a list of possible sets of values with their respective probabilities.

2.7. Expected Values in the Multivariate Case. Suppose the pair of chance variables X, Y has a joint probability distribution written in the general form of Sec. 2.6, and suppose $g(x, y)$ is a given function. Then the expected value of $g(X, Y)$, denoted by $E\{g(X, Y)\}$, is defined as the value $\sum_{i=1}^{k} \sum_{j=1}^{h} p_{ij}g(x_i,y_j)$. As a numerical example, suppose X, Y have the following joint probability distribution:

		X		
		-1	0	1
Y	2	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$
	4	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{6}$

Then, $E\{XY\} = (\frac{1}{12})(-1)(2) + (\frac{1}{4})(-1)(4) + (\frac{1}{12})(0)(2) + (\frac{1}{12})(0)(4) + (\frac{1}{3})(1)(2) + (\frac{1}{6})(1)(4) = \frac{1}{6}$.

The physical interpretation of $E\{g(X, Y)\}$ is exactly the same as in the case of a single chance variable. Given any joint probability distribution, we can construct a physical experiment defining a pair of chance variables with the given distribution. For example, this could be done by properly labeling the circumference of a round dial with pairs of numbers, and then spinning a well-balanced spinner mounted on the dial. The experiment can be performed N times, (t_i,μ_i) denoting the pair of numbers that appears on the ith performance of the experiment. Then if N is large, the average

$$\frac{g(t_1,\mu_1) + g(t_2,\mu_2) + \cdots + g(t_N,\mu_N)}{N}$$

can be expected to be close to $E\{g(X, Y)\}$.

Similarly, suppose we have a set of three chance variables X, Y, Z, with X having the possible values $x_1, \ldots, x_k$, Y having the possible values $y_1, \ldots, y_h$, Z having the possible values $z_1, \ldots, z_m$. Let p_{ijq} denote $P(X = x_i \text{ and } Y = y_j \text{ and } Z = z_q)$. Then, if $g(x,y,z)$ is a given function, the expected value of $g(X,Y,Z)$, denoted by $E\{g(X,Y,Z)\}$, is defined as

$$\sum_{i=1}^{k} \sum_{j=1}^{h} \sum_{q=1}^{m} p_{ijq} g(x_i, y_j, z_q)$$

The definition of expected value when we are dealing with more than three chance variables follows the same pattern.

If $r(x,y,z)$ and $s(x,y,z)$ are given functions and c, d are given constants, it follows directly from the definitions that $E\{cr(X,Y,Z) + ds(X,Y,Z)\} = c\,E\{r(X,Y,Z)\} + d\,E\{s(X,Y,Z)\}$.

2.8. Multivariate cdf's. If X, Y is a pair of chance variables with a given joint probability distribution, then for any given values x, y, we can compute $P(X \leq x \text{ and } Y \leq y)$. $P(X \leq x \text{ and } Y \leq y)$ is a function of x and y, and is called the "joint cumulative distribution function for X, Y." This joint cdf for X, Y contains the same information as the joint probability distribution in table form: The possible pairs of values are the points (x,y) where the function $P(X \leq x \text{ and } Y \leq y)$ has a jump, and the corresponding probability is the height of the jump.

Let us denote $P(X \leq x \text{ and } Y \leq y)$ by $F(x,y)$. If a_1, a_2, b_1, b_2 are any given values with $a_1 < a_2$ and $b_1 < b_2$, we shall prove that $P(a_1 < X \leq a_2 \text{ and } b_1 < Y \leq b_2) = F(a_2,b_2) - F(a_2,b_1) - F(a_1,b_2) + F(a_1,b_1)$. To prove this we define the events E_1, E_2, E_3, E_4 as follows:

E_1 is the event $(a_1 < X \leq a_2 \text{ and } b_1 < Y \leq b_2)$.
E_2 is the event $(X \leq a_1 \text{ and } b_1 < Y \leq b_2)$.
E_3 is the event $(a_1 < X \leq a_2 \text{ and } Y \leq b_1)$.
E_4 is the event $(X \leq a_1 \text{ and } Y \leq b_1)$.

The events E_1, E_2, E_3, E_4 are mutually exclusive by pairs. The event $(X \leq a_2 \text{ and } Y \leq b_2)$ is the same event as (E_1 or E_2 or E_3 or E_4); the event $(X \leq a_1 \text{ and } Y \leq b_2)$ is the same event as (E_2 or E_4); the event $(X \leq a_2 \text{ and } Y \leq b_1)$ is the same event as (E_3 or E_4). Thus we have

$F(a_2,b_2) = P(X \leq a_2 \text{ and } Y \leq b_2) = P(E_1 \text{ or } E_2 \text{ or } E_3 \text{ or } E_4) = P(E_1) + P(E_2) + P(E_3) + P(E_4)$; $F(a_1,b_2) = P(X \leq a_1 \text{ and } Y \leq b_2) = P(E_2 \text{ or } E_4) = P(E_2) + P(E_4)$; $F(a_2,b_1) = P(X \leq a_2 \text{ and } Y \leq b_1) = P(E_3 \text{ or } E_4) = P(E_3) + P(E_4)$; $F(a_1,b_1) = P(X \leq a_1 \text{ and } Y \leq b_1) = P(E_4)$

From these relationships, we find

$$P(a_1 < X \leq a_2 \text{ and } b_1 < Y \leq b_2) = F(a_2,b_2) - F(a_2,b_1) - F(a_1,b_2) + F(a_1,b_1)$$

2.9. Marginal Probability Distributions. Suppose the pair of chance variables X, Y has the joint probability distribution

$$\begin{array}{cc|cccc|}
 & & & X & & \\
 & & x_1 & x_2 & \cdots & x_k \\
\hline
 & y_1 & p_{11} & p_{21} & \cdots & p_{k1} \\
 & \cdot & \cdot & \cdot & & \cdot \\
Y & \cdot & \cdot & \cdot & & \cdot \\
 & \cdot & \cdot & \cdot & & \cdot \\
 & y_h & p_{1h} & p_{2h} & \cdots & p_{kh} \\
\hline
\end{array}$$

X by itself is a chance variable, with possible values $x_1, x_2, \ldots, x_k$. Also

$$P(X = x_i) = \sum_{j=1}^{h} p_{ij}$$

That is, we get $P(X = x_i)$ by adding all the probabilities in the column headed x_i in the table. To see this, we note that the event $(X = x_i)$ is the same event as $[(X = x_i$ and $Y = y_1)$ or $(X = x_i$ and $Y = y_2)$ or $\cdots$ or $(X = x_i$ and $Y = y_h)]$, and the events $(X = x_i$ and $Y = y_1)$, $(X = x_i$ and $Y = y_2)$, $\ldots$, $(X = x_i$ and $Y = y_h)$ are mutually exclusive by pairs; thus

$$P(X = x_i) = \sum_{j=1}^{h} P(X = x_i \text{ and } Y = y_j) = \sum_{j=1}^{h} p_{ij}$$

as stated.

Similarly, Y by itself is a chance variable, with possible values $y_1, y_2, \ldots, y_h$, and

$$P(Y = y_j) = \sum_{i=1}^{k} p_{ij}$$

by the same reasoning we used for X. These separate probability distributions of X and Y we get by adding probabilities from a joint probability distribution are called "marginal probability distributions." As a numerical example:

JOINT PROBABILITY DISTRIBUTION

$$\begin{array}{cc|ccc|}
 & & & X & \\
 & & -1 & 0 & 1 \\
\hline
Y & 2 & 1/12 & 1/12 & 1/3 \\
 & 4 & 1/4 & 1/12 & 1/6 \\
\hline
\end{array}$$

MARGINAL PROBABILITY DISTRIBUTION OF Y

Possible values	2	4
Probabilities	1/2	1/2

It should be noted that although a given joint probability distribution determines unique marginal probability distributions, it is not true that given marginal probability distributions determine a unique joint probability distribution. For example, the joint probability distribution

		X		
		-1	0	1
Y	2	$1/24$	$1/8$	$1/3$
	4	$7/24$	$1/24$	$1/6$

gives the same marginal probability distributions as the different joint probability distribution above.

Suppose X, Y have the joint cumulative distribution function $F(x,y)$. The chance variable X by itself has a cumulative distribution function, called the "marginal cumulative distribution function for X," denoted by $F_1(x)$. (The subscript 1 shows that it is the first of the two original variables we are dealing with.) We want to develop a formula that will enable us to compute $F_1(x)$ from a knowledge of $F(x,y)$. Since $F_1(x) = P(X < x)$, we have

$$F_1(x) = \sum_{i:x_i \leq x} \sum_{j=1}^{h} p_{ij} = F(x,\bar{y})$$

where $\bar{y}$ is any value greater than $\max(y_1, \ldots, y_h)$. We note that we may write $F(x,\bar{y})$ as

$$\lim_{y \to \infty} F(x,y)$$

which gives us the formula

$$F_1(x) = \lim_{y \to \infty} F(x,y)$$

Similarly, the marginal cdf for Y, $F_2(y)$ is given by the formula

$$F_2(y) = \lim_{x \to \infty} F(x,y)$$

If we have a set of three chance variables X, Y, Z, with X having possible values $x_1, \ldots, x_k$, Y having possible values $y_1, \ldots, y_h$, Z having possible values $z_1, \ldots, z_m$, let p_{ijq} denote $P(X = x_i$ and $Y = y_j$ and $Z = z_q)$. The joint marginal distribution of X, Y is given by

$$P(X = x_i \text{ and } Y = y_j) = \sum_{q=1}^{m} p_{ijq}$$

The marginal distribution of X is given by

$$P(X = x_i) = \sum_{j=1}^{h} \sum_{q=1}^{m} p_{ijq}$$

There are joint marginal distributions for X, Z and Y, Z and marginal distributions for Y and for Z, all derived from the joint probability distribution of X, Y, Z analogously.

The joint cdf for X, Y, Z is the function $F(x,y,z) = P(X \leqslant x$ and $Y \leqslant y$ and $Z \leqslant z)$. The joint marginal cdf for X, Y, denoted by $F_{1,2}(x,y)$, is given by

$$F_{1,2}(x,y) = \lim_{z\to\infty} F(x,y,z)$$

We also have

$$F_{1,3}(x,z) = \lim_{y\to\infty} F(x,y,z)$$

and

$$F_{2,3}(y,z) = \lim_{x\to\infty} F(x,y,z)$$

The marginal cdf for X is

$$F_1(x) = \lim_{y,z\to\infty} F(x,y,z)$$

with corresponding definitions for $F_2(y)$, $F_3(z)$.

2.10. Conditional Probability Distributions and Conditional Expected Values. Given the pair of chance variables X, Y, the "conditional probability distribution of X given that $Y = y_j$" is defined as the following probability distribution:

Possible values	x_1	x_2	$\cdots$	x_k
Probabilities	$P(X = x_1 \mid Y = y_j)$	$P(X = x_2 \mid Y = y_j)$	$\cdots$	$P(X = x_k \mid Y = y_j)$

where of course

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)}$$

in accordance with the definition of conditional probability given in Chap. 1. The "conditional expected value of $g(X)$ given that $Y = y_j$" is denoted by the symbol $E\{g(X) \mid Y = y_j\}$ and is defined as the quantity

$$\sum_{i=1}^{k} g(x_i)P(X = x_i \mid Y = y_j)$$

As a numerical example, if we have the joint distribution

$$\begin{array}{cc|ccc|} & & & X & \\ & & -1 & 0 & 1 \\ \hline Y & 2 & \tfrac{1}{12} & \tfrac{1}{12} & \tfrac{1}{3} \\ & 4 & \tfrac{1}{4} & \tfrac{1}{12} & \tfrac{1}{6} \\ \hline \end{array}$$

then the conditional distribution of X given that $Y = 2$ is

Possible values	-1	0	1
Probabilities	$\tfrac{1}{6}$	$\tfrac{1}{6}$	$\tfrac{2}{3}$

and $E\{X^2 \mid Y = 2\} = (\tfrac{1}{6})(-1)^2 + (\tfrac{1}{6})(0)^2 + (\tfrac{2}{3})(1)^2 = \tfrac{5}{6}$.

The physical interpretation of $E\{g(X) \mid Y = y_j\}$ is analogous to the interpretation of an ordinary expected value. We construct a physical experiment defining a pair of chance variables X, Y with the given joint probability distribution, and we perform the experiment N times. We disregard all outcomes except those on which the observed value of Y was y_j, and we take the average of the observed values of $g(X)$ on the outcomes where the observed value of Y was y_j. This average is expected to be close to $E\{g(X) \mid Y = y_j\}$.

If X, Y, Z are three chance variables with a given joint probability distribution, the joint conditional distribution of X, Y given that $Z = z_q$ is given by using the conditional probabilities $P(X = x_i$ and $Y = y_j \mid Z = z_q)$. The conditional probability distribution of X given that $Y = y_j$ and $Z = z_q$ is given by using the conditional probabilities $P(X = x_i \mid Y = y_j$ and $Z = z_q)$.

We can generalize the definitions above by using more general conditions. For example, we could define the conditional distribution of X given that $b < Y < c$ by using $P(X = x_i \mid b < Y < c)$.

Any conditional probability distribution has associated with it a conditional cdf. For example, the conditional cdf for X given that $Y = y_j$ is the function $P(X \leq x \mid Y = y_j)$.

2.11. Independent Chance Variables. X, Y are called independent if $P(X = x$ and $Y = y) = P(X = x)P(Y = y)$ for all values x, y. In terms of the tabled joint probability distribution, this means that the probability in any cell is the product of the two marginal probabilities in the row and column of that cell. Two examples illustrate this.

X, Y NOT INDEPENDENT

			X		
		−1	0	1	
Y	0	1/6	0	1/2	2/3
	1	0	1/3	0	1/3
		1/6	1/3	1/2	

X, Y INDEPENDENT

			X		
		−1	0	1	
Y	0	1/9	2/9	1/3	2/3
	1	1/18	1/9	1/6	1/3
		1/6	1/3	1/2	

If X, Y are independent, the conditional distribution of X given any condition on Y is the same as the marginal distribution of X. Also, the conditional distribution of Y given any condition on X is the same as the marginal distribution of Y.

If X, Y are independent and $r(x)$, $s(y)$ are any given functions, then

$$\begin{aligned} E\{r(X)s(Y)\} &= \sum_{j=1}^{h} \sum_{i=1}^{k} r(x_i)s(y_j)P(X = x_i \text{ and } Y = y_j) \\ &= \sum_{j=1}^{h} \sum_{i=1}^{k} r(x_i)s(y_j)P(X = x_i)P(Y = y_j) \\ &= \sum_{i=1}^{k} r(x_i)P(X = x_i) \sum_{j=1}^{h} s(y_j)P(Y = y_j) = E\{r(X)\}E\{s(Y)\} \end{aligned}$$

If X, Y are independent, then the joint cdf $F(x,y)$ is equal to $F_1(x)F_2(y)$ for all values x, y. To prove this, we have

$$\begin{aligned} F(x,y) &= \sum_{j:y_j \leq y} \sum_{i:x_i \leq x} P(X = x_i \text{ and } Y = y_j) = \sum_{j:y_j \leq y} \sum_{i:x_i \leq x} P(X = x_i)P(Y = y_j) \\ &= \sum_{i:x_i \leq x} P(X = x_i) \sum_{j:y_j \leq y} P(Y = y_j) = F_1(x)F_2(y) \end{aligned}$$

Conversely, since the joint cdf contains all the information about the joint distribution, if $F(x,y) = F_1(x)F_2(y)$ for all values x, y, then X and Y are independent.

The above statements can be generalized to more than two chance variables. X, Y, Z are independent if $P(X = x \text{ and } Y = y \text{ and } Z = z) = P(X = x)P(Y = y)P(Z = z)$ for all x, y, z. If X, Y, Z are independent, $E\{r(X)s(Y)t(Z)\} = E\{r(X)\}E\{s(Y)\}E\{t(Z)\}$. X, Y, Z are independent if and only if $F(x,y,z) = F_1(x)F_2(y)F_3(z)$ for all values x, y, z.

We can also define the independence of two sets of chance variables. If $X_1, \ldots, X_r, Y_1 \ldots, Y_s$ are chance variables such that $P(X_1 = x_1 \text{ and} \cdots \text{and } X_r = x_r \text{ and } Y_1 = y_1 \text{ and} \cdots \text{and } Y_s = y_s) = P(X_1 = x_1 \text{ and} \cdots \text{and } X_r = x_r)\ P(Y_1 = y_1 \text{ and} \cdots \text{and } Y_s = y_s)$, for all values $x_1, \ldots, x_r, y_1, \ldots, y_s$, then we say that the set $X_1, \ldots, X_r$ is independent of $Y_1, \ldots, Y_s$.

Whenever chance variables are defined by separated physical experiments, they are assumed to be independent. For example, if we shuffle two separate decks of numbered cards and define X as the number on the top of one deck and Y as the number on the top of the other deck, then X, Y are independent.

Chapter 3

CHANCE VARIABLES WITH AN INFINITE NUMBER OF POSSIBLE VALUES

3.1. Cases in Which the Possible Values Can Be Listed in an Infinite Sequence. We start with a simple example. Suppose our experiment is tossing a well-balanced coin until a head appears, and the chance variable X is defined as the total number of throws that must be made to have a head appear. In this case, the possible values of X are $1, 2, 3, \cdots$, and if r is a positive integer, $P(X = r) = P(\text{tail on toss 1 and tail on toss 2 and} \cdots \text{and tail on toss } r - 1 \text{ and head on toss } r) = P(\text{tail on toss 1})$ $P(\text{tail on toss 2}) \cdots P(\text{tail on toss } r - 1)\, P(\text{head on toss } r) = 1/2^r$. The probability distribution of X in table form is

Possible values	1	2	3	4	$\cdots$
Probabilities	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\cdots$

We note that the sum of the probabilities is the infinite series $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1$. The cdf $F(x)$ for this case is given as follows:

$$\begin{aligned}
F(x) &= 0 && \text{if } x < 1 \\
F(x) &= \tfrac{1}{2} && \text{if } 1 \leq x < 2 \\
F(x) &= \tfrac{1}{2} + \tfrac{1}{4} = \tfrac{3}{4} && \text{if } 2 \leq x < 3 \\
F(x) &= \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} = \tfrac{7}{8} && \text{if } 3 \leq x < 4
\end{aligned}$$

and so forth. In the general case of this sort, the probability distribution in table form will be represented by

Possible values	x_1	x_2	x_3	$\cdots$
Probabilities	p_1	p_2	p_3	$\cdots$

where the infinite series $p_1 + p_2 + p_3 + \cdots = 1$. $E\{g(X)\}$ is defined as the infinite series $p_1g(x_1) + p_2g(x_2) + p_3g(x_3) + \cdots$, provided that this series converges absolutely. If the series does not converge absolutely, we say that $E\{g(X)\}$ does not exist. Thus, for the numerical case given at the beginning of this section, $E\{X\} = (\frac{1}{2})(1) + (\frac{1}{4})(2) + (\frac{1}{8})(3) + \cdots = 2$; but $E\{2^X\} = (\frac{1}{2})(2^1) + (\frac{1}{4})(2^2) + (\frac{1}{8})(2^3) + \cdots = 1 + 1 + 1 + \cdots$, and this series does not converge, so $E\{2^X\}$ does not exist for the present case. When $E\{g(X)\}$ exists, its physical interpretation is exactly the same as in the case of a chance variable with a finite number of possible values. When $E\{g(X)\}$ fails to exist, it means that the average $[g(t_1) + \cdots + g(t_N)]/N$ (in the notation of Sec. 2.3) cannot be expected to be close to any finite number, because of the relatively frequent occurrence of quantities very large in absolute value among the numbers $g(t_1), \ldots, g(t_N)$.

3.2. Chance Variables with a Continuum of Possible Values. We introduce this case with a simple example. Suppose we have a well-balanced arrowheaded spinner mounted on a round dial with circumference equal to 1. The dial is labeled as in Fig. 3.1. We spin the spinner and define the chance variable X as the number on the dial to which the arrowhead will point (if the arrowhead points to the place that could be read as either 0 or 1, let us agree that it will be read as 1). In any such experiment that could actually be carried out in practice, it would be possible to read only a finite number of places on the dial, and thus X would have only a finite number of possible values. However, we are interested in the limiting ideal case where an infinite number of decimal places can be read, so that any value between 0 and 1 is a possible outcome of the experiment. We study this case by finding the cdf. If we could read only two decimal places, the cdf for X would have 100 equally spaced jumps, each of height $\frac{1}{100}$. If we could read three decimal places, the cdf would have 1,000 equally spaced jumps, each of height $\frac{1}{1000}$. Clearly, as the number of decimal places we can read increases, the cdf approaches the function $F(x)$ defined as follows: $F(x) = 0$ for $x \leqslant 0$, $F(x) = x$ for $0 \leqslant x \leqslant 1$, $F(x) = 1$ for $x \geqslant 1$. This function $F(x)$ is the cdf for X in the limiting case where an infinite number of decimal places can be read on the dial. The fact that $P(X \leqslant x) = x$ for

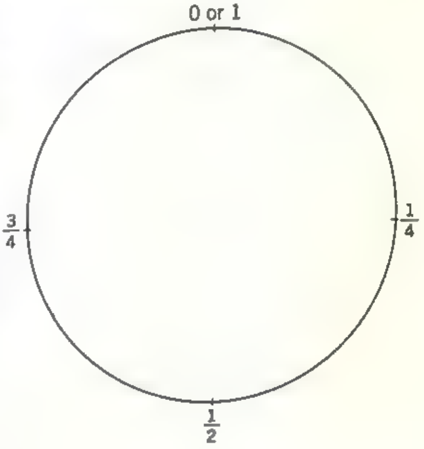


Fig. 3.1

$0 \leq x \leq 1$ in the limiting case is also obvious from the intuitive meaning of "well-balanced spinner." For example, $P(X \leq \frac{1}{4})$ must be equal to $\frac{1}{4}$, since the set of points on the dial which are labeled with values no greater than $\frac{1}{4}$ cover one-quarter of the total circumference.

3.3. General Properties of cdf's. In Sec. 3.2 we discussed a cdf that was a continuous function. Earlier, we had discussed cdf's that had discontinuities. The following properties are possessed by every cdf $F(x)$:

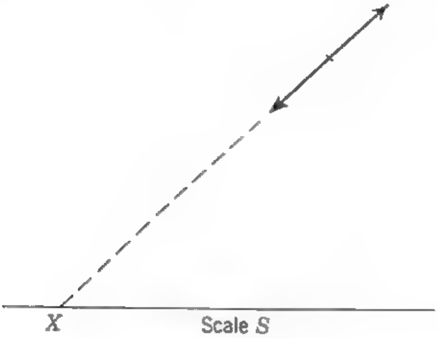


Fig. 3.2

1. For any values a, b, with $a < b$, $F(a) \leq F(b)$.
2. $\lim_{x \to -\infty} F(x) = 0$.
3. $\lim_{x \to \infty} F(x) = 1$.
4. $\lim_{\substack{\Delta \to 0 \\ (\Delta > 0)}} [F(x + \Delta) - F(x)] = 0$,

for every value x.

Conversely, any function that possesses these four properties is the cdf for some chance variable X. The following type of physical experiment can be used to define chance variables with a great variety of cdf's. A well-balanced spinner, with an arrowhead at each end, is set spinning. Below the spinner is a line extending indefinitely in both directions, and on the line is engraved a scale S, which can be chosen arbitrarily. When the spinner comes to rest, an imaginary line is drawn through the spinner and extended until it meets the scale S (the probability is 0 that the spinner will be parallel to S). The chance variable X is the value on S intersected by the imaginary line through the spinner. Figure 3.2 shows this.

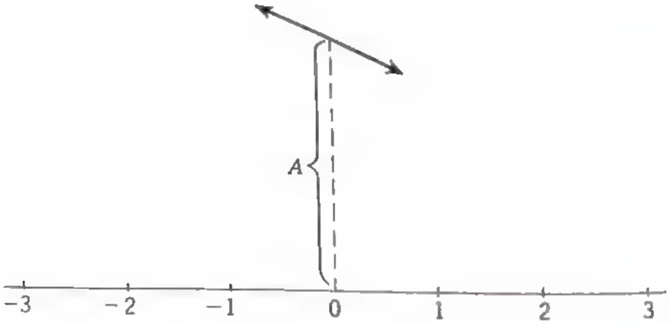


Fig. 3.3

The cdf for X is determined by the scale S. Some examples follow.

Example 1. The scale S is an ordinary arithmetic scale, and the center of the spinner is A units vertically above the zero point on the scale S, as in Fig. 3.3. Remembering that the spinner is well balanced, it is easily found from the definition of the angle θ in Fig. 3.4 that $F(x) = P(X \leq x) = \theta/\pi = \frac{1}{2} - 1/\pi \arctan x/A$. The graph of $F(x)$ in this case is shown in Fig. 3.5.

Example 2. The scale S is broken into subintervals, and every point in a given subinterval is labeled with the same number, as in Fig. 3.6. In such a case, if the scale S is broken into a finite number of subintervals (so that the subintervals on each end are of infinite length), then X has only a finite number of possible values, and the cdf $F(x)$ is of the type discussed in Chap. 2: it is horizontal except at the finite number of points at which it has a jump. If the scale S is broken into an infinite number of subintervals, then the chance variable X has an infinite number of possible values, which can be listed in a sequence, as in Sec. 3.1. The cdf in this case is horizontal except at the infinite number of points where it has a jump.

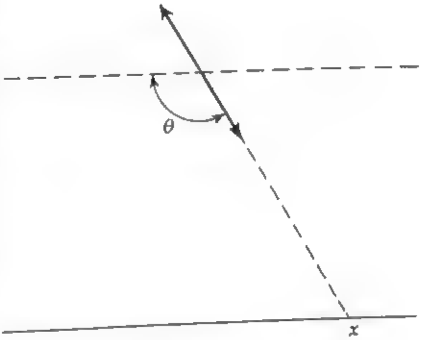


Fig. 3.4

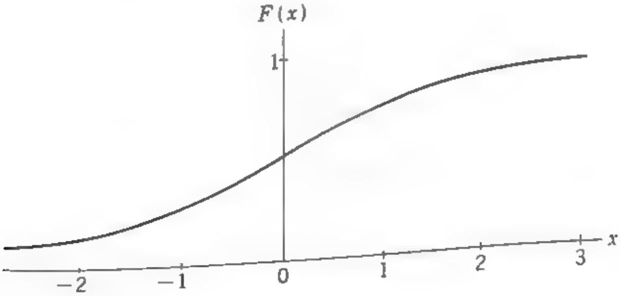


Fig. 3.5

Example 3. This example is a mixture of Examples 1 and 2. The scale S is broken into subintervals, and some subintervals have all their points labeled with the same number (as in Example 2), while the other subintervals are labeled with arithmetic scales. Then the cdf will have the type of graph shown in Fig. 3.7.

No matter what type of cdf $F(x)$ the chance variable X has, we see from

Sec. 2.4 that for any given values a, b, with $a < b$, we have $P(a < X \leq b) = F(b) - F(a)$. If we hold a fixed and let b approach a, we know from property 4 of a cdf that $P(a < X \leq b)$ approaches zero. However, if we fix b and let a approach b, then $P(a < X \leq b)$ approaches the jump in $F(x)$ at the point b [if $F(x)$ is continuous at the point b, it has a jump of 0]. But it is clear that

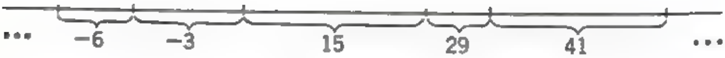


Fig. 3.6

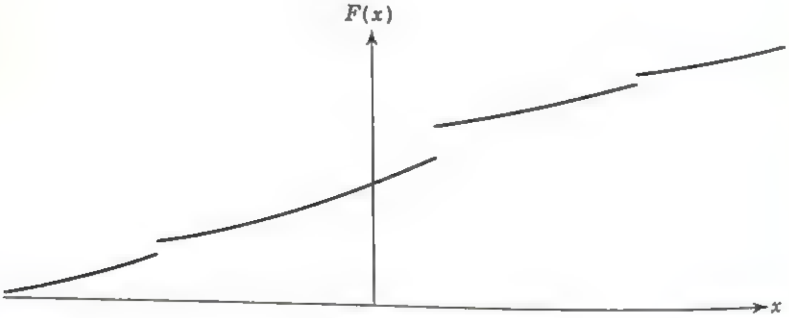


Fig. 3.7

$$\lim_{a \to b} P(a < X \leq b) = P(X = b)$$

Thus if a cdf is continuous at a point b, $P(X = b) = 0$.

3.4. Probability Density Functions. Suppose the cdf $F(x)$ is continuous everywhere and has a derivative at all points, except perhaps a finite number of exceptional points. Then the derivative is called the "probability density function," or "pdf," for the chance variable X. As a matter of notation, the derivatives of the cdf's $F(x)$, $G(x)$ will be denoted by $f(x)$, $g(x)$, respectively.

As an example, the pdf $f(x)$ corresponding to the cdf $F(x)$ of Sec. 3.2 is

$f(x) = 1$ for $0 < x < 1$; $f(x) = 0$ for $x < 0$ or $x > 1$; $f(x)$ does not exist at $x = 0$ or $x = 1$.

As another example, the pdf $f(x)$ corresponding to the cdf $F(x)$ of Example 1 of Sec. 3.3 is

$$f(x) = \frac{1}{\pi A} \frac{1}{1 + (x/A)^2}$$

If $F(x)$ is a cdf and $f(x)$ is the corresponding pdf, then

$$\int_a^b f(x)\,dx = F(x)\Big]_a^b = F(b) - F(a).$$

In particular,

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1 - 0 = 1$$

Thus any pdf has the following two properties:

1. $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1.$

2. $f(x) \geqslant 0$ for each x, since $f(x)$ is the derivative of $F(x)$, a nondecreasing function.

Conversely, any function with these two properties is the pdf for some chance variable X.

Since

$$\int_{-\infty}^{b} f(x)\,dx = F(b) - \lim_{x\to-\infty} F(x) = F(b)$$

if we are given a pdf, we can find the corresponding cdf by integrating.

3.5. Expected Values. If X is a chance variable with pdf $f(x)$ and $g(x)$ is a given function, then $E\{g(X)\}$ is defined as $\int_{-\infty}^{\infty} g(x)f(x)\,dx$.

The physical interpretation of $E\{g(X)\}$ is the same as in the case where X has only a finite number of possible values. To see this, suppose X has only a finite number of possible values, say, $x_1, x_2, \ldots, x_k$, with probabilities $p_1, p_2, \ldots, p_k$. Suppose A is a value less than the smallest of the values $x_1, x_2, \ldots, x_k$. Then

$$\begin{aligned} E\{g(X)\} &= \sum_{i=1}^{k} g(x_i)p_i \\ &= \lim_{\Delta x\to 0} \sum_{j=0}^{\infty} g(A + j\,\Delta x)[F(A + (j+1)\,\Delta x) - F(A + j\,\Delta x)] \end{aligned}$$

because $F[A + (j+1)\,\Delta x] - F(A + j\,\Delta x) = 0$, unless one or more of the values x_i falls between $A + j\,\Delta x$ and $A + (j+1)\,\Delta x$, and if only one of the values x_i falls in the interval, $F[A + (j + 1)\,\Delta x] - F(A + j\,\Delta x) = p_i$. Now if we return to the case where X has a pdf $f(x)$, we find

$$\lim_{\Delta x \to 0} \sum_{j=0}^{\infty} g(A + j\,\Delta x)[F(A + (j + 1)\,\Delta x) - F(A + j\,\Delta x)] = \int_A^{\infty} g(x) f(x)\, dx$$

Letting A approach $-\infty$, we get $E\{g(X)\}$.

As an example of the computation of an expected value, suppose X has pdf $f(x)$ defined as follows: $f(x) = 1$ if $0 < x < 1$; $f(x) = 0$ for other values of x. Then

$$E\{X\} = \int_{-\infty}^{\infty} xf(x)\, dx = \int_{-\infty}^{0} xf(x)\, dx + \int_0^1 xf(x)\, dx + \int_1^{\infty} xf(x)\, dx$$

$$= \int_{-\infty}^{0} x0\, dx + \int_0^1 x1\, dx + \int_1^{\infty} x0\, dx = 0 + \tfrac{1}{2} + 0 = \tfrac{1}{2}$$

Also

$$E\{X^2\} = \int_{-\infty}^{\infty} x^2 f(x)\, dx = 0 + \tfrac{1}{3} + 0 = \tfrac{1}{3}$$

3.6. Expected Values in More Complicated Cases. We now have formulas for $E\{g(X)\}$ in cases where the cdf $F(x)$ increases only in jumps and in cases where $F(x)$ is continuous and has a derivative. But Example 3 of Sec. 3.3 showed that some cdf's are mixtures, increasing both continuously and in jumps. If $F(x)$ is such a mixed cdf, we have $F(x) = R(x) + S(x)$, where $R(x)$ increases only in jumps and $S(x)$ is a continuous function. $R(x)$ and $S(x)$ have all the properties of cdf's, except that

$$\lim_{x \to \infty} R(x) < 1 \qquad \text{and} \qquad \lim_{x \to \infty} S(x) < 1$$

since

$$\lim_{x \to \infty} [R(x) + S(x)] = 1$$

Let $x_1, x_2, \ldots,$ denote the points on the x axis at which $R(x)$ has jumps, and let $r(x_i)$ denote the height of the jump in $R(x)$ at x_i. Also, we assume that $S(x)$ has a derivative $s(x)$ everywhere, except perhaps at a finite number of points. Then $E\{g(X)\}$ is defined as $\sum_i g(x_i)r(x_i) + \int_{-\infty}^{\infty} g(x)s(x)\, dx$. This has the usual physical interpretation.

3.7. Transformation of a Chance Variable. Sometimes we start with a chance variable X, but find it more convenient to deal with another chance variable defined as a given function of X. For example, if X is the distance from the point of impact of a missile to the target, it might turn out that with respect to the effect on the target, X^2 is the important quantity. Then, instead of the probability distribution of X, we should be interested in the probability distribution of X^2. And in general, we might be interested in the probability distribution of the chance variable $h(X)$ [where $h(x)$ is some given function] rather than the probability distribution of X itself. We note that $h(X)$ is a chance variable which can be defined by the same physical experiment that defines X: the experiment can be performed, and X observed, but then replaced by $h(X)$. Thus we are merely relabeling our outcome.

We should like to develop methods for finding the probability distribution of $h(X)$, starting from the known probability distribution of X. In a case where X has only a finite number of possible values, so that the probability distribution can be given in table form, it is a very simple matter to find the probability distribution of $h(X)$ by relabeling the entries in the table. Two examples will make this clear.

Example 1. If the probability distribution of X is

-3	0	5
$1/2$	$1/6$	$1/3$

then the probability distribution of $2X + 3$ is

$-3 = 2(-3) + 3$	$3 = 2(0) + 3$	$13 = 2(5) + 3$
$1/2$	$1/6$	$1/3$

Example 2. [In this example, two different values of X lead to the same value of $h(X)$, so the probabilities must be added together.] If the probability distribution of X is

-3	-2	0	1	2
$1/5$	$1/5$	$1/5$	$1/5$	$1/5$

then the probability distribution of X^2 is

0	1	4	9
$1/5$	$1/5$	$2/5$	$1/5$

In a case where X has an infinite number of possible values which can be listed in a sequence, so that the distribution can be given in table form, the situation is the same as in the examples just discussed. Thus, if the probability distribution of X is

1	2	3	4	$\cdots$
$1/4$	$1/2$	$1/8$	$1/16$	$\cdots$

then the probability distribution of $2X$ is

2	4	6	8	$\cdots$
½	¼	⅛	1⁄16	$\cdots$

while the probability distribution of $(X^2 - 1)$ is

0	3	8	15	$\cdots$
½	¼	⅛	1⁄16	$\cdots$

In many important problems, it is necessary to find the distribution of $h(X)$ when X has a continuous cdf $F(x)$ with corresponding pdf $f(x)$. In such a case, the distribution of X cannot be given in the form of a table, since the possible values of X fill a continuum, so new techniques will be required to find the probability distribution of $h(X)$. We assume that $h(x)$ has a derivative at every x. Also, for the time being, we assume that $h(x)$ is either a strictly increasing function of x or else a strictly decreasing function of x. For convenience, we denote the chance variable $h(X)$ by Y, and we denote the cdf for Y by $K(y)$; that is, $P(Y \leqslant y) = K(y)$. Also, we denote by $r(Y)$ the function that X is of Y. For example, if $h(X) = 2X + 3$, then $Y = 2X + 3$, so $X = (Y - 3)/2$ and $r(Y) = (Y - 3)/2$. For a given value y, the event $(Y \leqslant y)$ is the same event as $[h(X) \leqslant y]$, which in turn is the same event as $[X \leqslant r(y)]$ if $h(x)$ is an increasing function, or is the same event as $[X \geqslant r(y)]$ if $h(x)$ is a decreasing function. From this it follows that if $h(x)$ is increasing, $K(y) = F[r(y)]$; if $h(x)$ is decreasing, $K(y) = 1 - F[r(y)]$. Then, by the rules for differentiation, $(d/dy)K(y)$ exists and is equal to $f[r(y)](|\,dr(y)/dy\,|)$, whether $r(y)$ is increasing or decreasing. Thus the pdf for the chance variable $Y = h(X)$ is equal to $f[r(y)]\,|\,dr(y)/dy\,|$.

A second method of finding the pdf for Y will now be described. If $k(y)$ is the pdf for Y and $g(y)$ is a given function, then

$$E\{g(Y)\} = \int_{-\infty}^{\infty} g(y)k(y)\,dy$$

But

$$g(Y) = g[h(X)] \qquad \text{and} \qquad E\{g[h(X)]\} = \int_{-\infty}^{\infty} g[h(x)]f(x)\,dx$$

In this last integral, we make the change of variable $y = h(x)$, the integral becoming

$$\int_{-\infty}^{\infty} g(y)f[r(y)]\left|\frac{dr(y)}{dy}\right| dy$$

Thus we have

$$\int_{-\infty}^{\infty} g(y)k(y)\,dy = \int_{-\infty}^{\infty} g(y)f[r(y)]\left|\frac{dr(y)}{dy}\right| dy$$

for each function $g(y)$. The only way for this to hold is to have

$$k(y) = f[r(y)]\left|\frac{dr(y)}{dy}\right|$$

which is what we found above. Two examples illustrate the discussion.

Example 1. X has the pdf $f(x)$ defined as follows: $f(x) = 0$ for $x < 0$; $f(x) = 1$ for $0 \leqslant x \leqslant 1$; $f(x) = 0$ for $x > 1$. $Y = 2X - 3$. Then $X = (Y + 3)/2 = r(Y)$, so $dr(y)/dy = \frac{1}{2}$; and $f[r(y)] = 1$ for $0 \leqslant (y + 3)/2 \leqslant 1$, or $-3 \leqslant y \leqslant -1$, and $f[r(y)] = 0$ if $y < -3$ or $y > -1$. Thus $k(y)$, the pdf for Y, is given as follows: $k(y) = 0$ if $y < -3$; $k(y) = \frac{1}{2}$ if $-3 \leqslant y \leqslant -1$; $k(y) = 0$ if $y > -1$.

Example 2. X has the pdf $f(x) = e^{-x}$ for $x \geqslant 0$, $f(x) = 0$ for $x < 0$. $Y = X^3$, so $X = Y^{1/3} = r(Y)$, and $dr(y)/dy = (1/3y^{2/3})$. $f[r(y)] = 0$ for $y^{1/3} < 0$, and $f[r(y)] = e^{-y^{1/3}}$ for $y^{1/3} \geqslant 0$. Thus $k(y) = 0$ for $y < 0$, $k(y) = (1/3y^{2/3})\, e^{-y^{1/3}}$ for $y > 0$.

Our discussion so far has assumed that Y is a monotonic function of X. As an example of a case where this is not so, suppose $Y = X^2$ and we want to find $k(y)$, the pdf for Y. Denoting the cdf for Y by $K(y)$, it is clear that $K(y) = 0$ for $y \leqslant 0$. The event $(Y \leqslant y)$ is the same event as $(X^2 \leqslant y)$, and if $y \geqslant 0$, this is the same event as $(-\sqrt{y} \leqslant X \leqslant \sqrt{y})$. Therefore if $y > 0$, $P(Y \leqslant y) = P(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y}) = K(y)$, and $k(y) = (d/dy)K(y) = (1/2\sqrt{y})f(\sqrt{y}) + (1/2\sqrt{y}) \times f(-\sqrt{y})$. For example, if $f(x) = \dfrac{1}{\pi}\dfrac{1}{1 + x^2}$ for all x, then $k(y) = 0$ for $y < 0$; $k(y) = \dfrac{1}{\pi\sqrt{y}}\dfrac{1}{1 + y}$ for $y > 0$; $k(y)$ does not exist at $y = 0$.

3.8. Pairs of Chance Variables. We introduced the joint cdf $F(x,y)$ for a pair of chance variables in Chap. 2. Now suppose we are dealing with a case where the possible pairs of values of (X, Y) fill a continuum in two-dimensional space, and $F(x,y)$ is continuous everywhere, and the second derivative $\partial^2F(x,y)/\partial x\, \partial y$ exists, except perhaps at points lying on a finite number of curves in the plane. In such a case, $\partial^2F(x,y)/\partial x\, \partial y$ is called "the joint pdf for X, Y" and will be denoted by $f(x,y)$.

If a_1, a_2, b_1, b_2 are any given values with $a_1 < a_2$ and $b_1 < b_2$, then

$$\int_{b_1}^{b_2}\int_{a_1}^{a_2} f(x,y)\,dx\,dy = \int_{b_1}^{b_2}\left[\frac{\partial F(a_2,y)}{\partial y} - \frac{\partial F(a_1,y)}{\partial y}\right]dy$$
$$= F(a_2,b_2) - F(a_2,b_1) - F(a_1,b_2) + F(a_1,b_1)$$

But in Sec. 2.8, we proved that this last expression is equal to $P(a_1 < X \leqslant a_2$ and $b_1 < Y \leqslant b_2)$. Thus we find

$$\int_{b_1}^{b_2}\int_{a_1}^{a_2} f(x,y)\,dx\,dy = P(a_1 < X \leqslant a_2 \text{ and } b_1 < Y \leqslant b_2)$$

From Sec. 2.9, we have that

$$P(X \leqslant x) = \lim_{y\to\infty} F(x,y)$$

and this gives

$$\lim_{x\to-\infty}\lim_{y\to\infty} F(x,y) = \lim_{x\to-\infty} P(X \leqslant x) = 0$$

and

$$\lim_{x\to\infty}\lim_{y\to\infty} F(x,y) = \lim_{x\to\infty} P(X \leqslant x) = 1$$

Similarly, we find

$$\lim_{y\to-\infty}\lim_{x\to\infty} F(x,y) = 0$$

From now on, we shall denote $\lim\limits_{x\to\infty}\lim\limits_{y\to-\infty} F(x,y)$ by $F(\infty,-\infty)$, etc. We have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = F(\infty,\infty) - F(\infty,-\infty) - F(-\infty,\infty) + F(-\infty,-\infty) = 1 - 0 - 0 + 0 = 1$$

$f(x,y)$ cannot be negative at any point x, y, for if it were, we could find a small rectangle ($a_1 < x \leqslant a_2$, $b_1 < y \leqslant b_2$) around that point, throughout which $f(x,y)$ is negative. But then $\int_{b_1}^{b_2}\int_{a_1}^{a_2} f(x,y)\,dx\,dy$ would be negative, which is impossible, since we saw above that this integral gives $P(a_1 < X \leqslant a_2$ and $b_1 < Y \leqslant b_2)$.

Thus we have shown that any joint pdf $f(x,y)$ has the following two properties:

1. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$
2. $f(x,y) \geqslant 0$ at every point x, y.

Conversely, any function $f(x,y)$ with these two properties is the joint pdf for some pair of chance variables X, Y.

$P(X \leqslant x$ and $Y \leqslant y)$ can be written as $P(X \leqslant x)P(Y \leqslant y \mid X \leqslant x)$, and we know from above that

$$\lim_{x\to-\infty} P(X \leqslant x) = 0$$

Therefore

$$\lim_{x\to-\infty} P(X \leqslant x \text{ and } Y \leqslant y) = F(-\infty,y) = 0 \qquad \text{for any } y$$

Similarly,

$$F(x,-\infty) = 0 \qquad \text{for any } x$$

Since

$$\int_{-\infty}^{y} \int_{-\infty}^{x} f(x,y)\,dx\,dy = F(x,y) - F(x,-\infty) - F(-\infty,y) + F(-\infty,-\infty)$$
$$= F(x,y) - 0 - 0 + 0 = F(x,y)$$

if we are given $f(x,y)$, we can find $F(x,y)$ by integrating.

If X, Y have the joint pdf $f(x,y)$ and if S is a set of points in the plane, then

$$P(X,\ Y \text{ in } S) = \iint_S f(x,y)\,dx\,dy$$

This has already been shown for the case where S is a rectangle. In the general case, S consists of nonoverlapping rectangles and parts of rectangles and $P(X,\ Y \text{ in } S)$ is the sum of the probabilities assigned to these nonoverlapping parts of S. By using the standard argument involving breaking S into finer and finer pieces, we can show $P(X,\ Y$ in $S) = \iint_S f(x,y)\,dx\,dy$.

If X, Y have joint pdf $f(x,y)$ and if $g(x,y)$ is a given function, then

$$E\{g(X,Y)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\,dx\,dy$$

if this integral exists. The physical interpretation of $E\{g(X,Y)\}$ is the usual one. As an example of the computation, if $f(x,y) = e^{-(x+y)}$ for $x > 0$ and $y > 0$, $f(x,y) = 0$ if $x \leq 0$ or $y \leq 0$, then

$$E\{X\,Y\} = \int_0^{\infty} \int_0^{\infty} xye^{-(x+y)}\,dx\,dy = 1$$

3.9. Transformation of Chance Variables. Just as in the case of a single chance variable, sometimes we start with a pair of chance variables X, Y but find it desirable to deal with a new pair of chance variables defined as given functions of X and Y. Then the problem is to find the joint distribution of the new pair of chance variables. In a case where X, Y have only a finite number of possible pairs of values, we can simply relabel the possible values to find the joint distribution of the new pair of chance variables.

A more complicated situation is where X, Y have a joint pdf $f(x,y)$. Suppose we want to find the joint distribution of the pair of chance variables W, Z, where $W = r(X,Y), Z = s(X,Y)$, the functions $r(x,y)$ and $s(x,y)$ having continuous first partial derivatives everywhere and also having the property that for any given values w and z, the simultaneous equations $w = r(x,y)$ and $z = s(x,y)$ have exactly one solution in x and y.

Then the equations $w = r(x,y)$ and $z = s(x,y)$ can be solved for x and y in terms of w and z, to give, say, $x = t(w,z)$ and $y = u(w,z)$. We denote by $J(w,z)$ the determinant

$$\begin{vmatrix} \dfrac{\partial t(w,z)}{\partial w} & \dfrac{\partial t(w,z)}{\partial z} \\ \dfrac{\partial u(w,z)}{\partial w} & \dfrac{\partial u(w,z)}{\partial z} \end{vmatrix}$$

which is a function of w and z. We want to find the joint pdf for W, Z. Denote this joint pdf by $k(w,z)$. For a given function $g(w,z)$, we have

$$E\{g(W,Z)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(w,z)k(w,z)\, dw\, dz$$

But $W = r(X,Y)$ and $Z = s(X,Y)$, so that

$$E\{g(W,Z)\} = E\{g[r(X,Y), s(X,Y)]\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g[r(x,y), s(x,y)]f(x,y)\, dx\, dy$$

In this second integral we make the change of variable $w = r(x,y)$, $z = s(x,y)$, and the integral becomes

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(w,z)f[t(w,z), u(w,z)]\, |J(w,z)|\, dw\, dz$$

which must equal $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(w,z)k(w,z)\, dw\, dz$ for every function $g(w,z)$. The only way for this to happen is to have $k(w,z) = f[t(w,z), u(w,z)]\, |J(w,z)|$. This tells us what the pdf $k(w,z)$ is.

Two examples illustrate the discussion.

Example 1. $f(x,y) = (1/2\pi) \exp[-(1/2)(x^2 + y^2)]$ for all x, y. $W = X^2Y^3$, $Z = X^3Y^2$. Then we find $X = Z^{3/5}W^{-2/5}$, $Y = W^{3/5}Z^{-2/5}$, so $t(w,z) = z^{3/5}w^{-2/5}$ and $u(w,z) = w^{3/5}z^{-2/5}$. Then the determinant $J(w,z)$ is

$$\begin{vmatrix} -(2/5)z^{3/5}w^{-7/5} & (3/5)z^{-2/5}w^{-2/5} \\ (3/5)z^{-2/5}w^{-2/5} & -(2/5)w^{3/5}z^{-7/5} \end{vmatrix} = -(1/5)(wz)^{-4/5}$$

Finally, $k(w,z) = (wz^{-4/5}/10\pi) \exp[-(1/2)(z^{6/5}w^{-4/5} + z^{-4/5}w^{6/5})]$.

Example 2. $f(x,y) = e^{-(x+y)}$ if $x > 0$ and $y > 0$; $f(x,y) = 0$ if either $x \leqslant 0$ or $y \leqslant 0$. $W = 2X + Y$, $Z = X + 3Y$. Then we find $X = (3/5)W - (1/5)Z$, $Y = -(1/5)W + (2/5)Z$, so $t(w,z) = (3/5)w - (1/5)z$ and $u(w,z) = -(1/5)w + (2/5)z$. Then the determinant $J(w,z)$ is

$$\begin{vmatrix} 3/5 & -1/5 \\ -1/5 & 2/5 \end{vmatrix} = 1/5$$

Finally, $k(w,z) = (\frac{1}{5}) \exp[-((\frac{3}{5})w - (\frac{1}{5})z - (\frac{1}{5})w + (\frac{2}{5})z)]$ if $(\frac{3}{5})w - (\frac{1}{5})z > 0$ and $-(\frac{1}{5})w + (\frac{2}{5})z > 0$, while $k(w,z) = 0$ for other values of w, z.

3.10. Marginal and Conditional pdf's. If X, Y have the joint pdf $f(x,y)$, we can write the marginal cdf $F_1(x)$ as $\int_{-\infty}^{x}\int_{-\infty}^{\infty} f(r,y)\,dy\,dr$, and then we have

$$\frac{dF_1(x)}{dx} = \int_{-\infty}^{\infty} f(x,y)\,dy$$

Thus the marginal pdf for X, usually denoted by $f_1(x)$, is given by $\int_{-\infty}^{\infty} f(x,y)\,dy$. Similarly, $f_2(y)$, the marginal pdf for Y, is given by $\int_{-\infty}^{\infty} f(x,y)\,dx$.

Suppose that R is a set of points on the y axis with $\int_R f_2(y)\,dy > 0$. Then the conditional cdf for X given that Y is in R, $F_1(x \mid Y \text{ in } R)$, is equal to

$$P(X \leq x \mid Y \text{ in } R) = \frac{P(X \leq x \text{ and } Y \text{ in } R)}{P(Y \text{ in } R)} = \frac{\int_{-\infty}^{x}\int_R f(r,y)\,dy\,dr}{\int_R f_2(y)\,dy}$$

Then

$$\frac{dF_1(x \mid Y \text{ in } R)}{dx} = \frac{\int_R f(x,y)\,dy}{\int_R f_2(y)\,dy}$$

which gives the conditional pdf for X, given that Y is in R, denoted by $f_1(x \mid Y \text{ in } R)$. The conditional expected value of $g(X)$ given that Y is in R is defined as $\int_{-\infty}^{\infty} g(x) f_1(x \mid Y \text{ in } R)\,dx$ and is denoted by $E\{g(X) \mid Y \text{ in } R\}$. If the set R is taken as the interval $(y, y + \Delta y)$, we find

$$\lim_{\Delta y \to 0} f_1(x \mid y \leq Y \leq y + \Delta y) = \frac{f(x,y)}{f_2(y)}$$

and this last expression is called "the conditional pdf for X given that $Y = y$" and is denoted by $f_1(x \mid Y = y)$. Then $E\{g(X) \mid Y = y\}$ is defined as

$$\int_{-\infty}^{\infty} g(x) \frac{f(x,y)}{f_2(y)}\,dx$$

The chance variables X, Y were defined to be independent if $F(x,y) = F_1(x)F_2(y)$ for all values x, y. If X, Y have joint pdf $f(x,y)$ and X, Y are independent, we have

$$F(x,y) = F_1(x)F_2(y)$$

$$\frac{\partial F(x,y)}{\partial y} = F_1(x)f_2(y)$$

$$\frac{\partial^2 F(x,y)}{\partial x\,\partial y} = f_1(x)f_2(y)$$

so $f(x,y) = f_1(x)f_2(y)$ for all x, y. Conversely, if $f(x,y) = f_1(x)f_2(y)$ for all x, y, we have

$$\begin{aligned} F(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f(r,s)\,dr\,ds \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_1(r)f_2(s)\,dr\,ds \\ &= \left[\int_{-\infty}^{x} f_1(r)\,dr\right]\left[\int_{-\infty}^{y} f_2(s)\,ds\right] \\ &= F_1(x)F_2(y) \end{aligned}$$

so that X, Y are independent if and only if $f(x,y) = f_1(x)f_2(y)$ for all x, y. It is easily verified that if X, Y are independent, then $f_1(x \mid Y \text{ in } R) = f_1(x)$, $f_1(x \mid Y = y) = f_1(x)$, $E\{g(X) \mid Y \text{ in } R\} = E\{g(X)\}$, with similar statements for $f_2(y \mid X \text{ in } S)$, etc.

As an example of a joint distribution for two chance variables which are independent, suppose $f(x,y) = 1$ when $0 < x < 1$ and $0 < y < 1$, while $f(x,y) = 0$ if $x \leq 0$ or $x \geq 1$ or $y \leq 0$ or $y \geq 1$. Then we find that if $0 < x < 1$,

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 f(x,y)\,dy = \int_0^1 1\,dy = 1$$

while if $x \leq 0$ or $x \geq 1$,

$$f_1(x) = \int_{-\infty}^{\infty} 0\,dy = 0$$

Similarly, $f_2(y) = 1$ for $0 < y < 1$, $f_2(y) = 0$ for $y \leq 0$ or $y \geq 1$. Thus we have $f(x,y) = f_1(x)f_2(y)$ for all x, y.

As an example of a joint distribution for two chance variables which are not independent, suppose $f(x,y) = 2$ when $0 < x$ and $0 < y$ and $x + y < 1$, while $f(x,y) = 0$ otherwise. From this, we find that $f_1(x) = 2(1 - x)$ for $0 < x < 1$, $f_1(x) = 0$ otherwise. Also, $f_2(y) = 2(1 - y)$ for $0 < y < 1$, $f_2(y) = 0$ otherwise. Thus $f(x,y)$ is not equal to $f_1(x)f_2(y)$ for all values x, y.

3.11. More Than Two Jointly Distributed Chance Variables. If $X_1, X_2, \ldots, X_n$ is a set of n jointly distributed chance variables, their joint cdf $F(x_1, x_2, \ldots, x_n)$ is defined as $P(X_1 \leq x_1$ and $\cdots$ and $X_n \leq x_n)$. If any of the n quantities $x_1, \ldots, x_n$ is $-\infty$, then $F(x_1, \ldots, x_n) = 0$. $F(\infty, \ldots, \infty) = 1$. The marginal cdf for X_1, X_2, X_3, for example, is given by $F(x_1, x_2, x_3, \infty, \ldots, \infty)$ and is denoted by $F_{1,2,3}(x_1,x_2,x_3)$. The marginal cdf for X_i, denoted by $F_i(x_i)$, is given by $F(\infty, \infty, \ldots, \infty, x_i, \infty, \ldots, \infty)$. $X_1, X_2, \ldots, X_n$ are independent if $F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n)$ for all values of $x_1, x_2, \ldots, x_n$.

If $F(x_1, x_2, \ldots, x_n)$ is continuous everywhere, and

$$\frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1\, \partial x_2 \cdots \partial x_n}$$

exists everywhere, with the possible exception of a finite number of curves in n-dimensional space, this derivative is called the joint pdf for $X_1, \ldots, X_n$ and is denoted by $f(x_1, \ldots, x_n)$. If R is a set of points in n-dimensional space, then

$$P[(X_1, \ldots, X_n) \text{ in } R] = \int \cdots \int_R f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

$E\{g(X_1, \ldots, X_n)\}$ is defined as $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$. If new chance variables $Y_1, \ldots, Y_n$ are defined by $Y_i = r_i(X_1, \ldots, X_n)$, where the functions $r_1(x_1, \ldots, x_n), \ldots, r_n(x_1, \ldots x_n)$ have continuous first partial derivatives and allow a unique solution for $X_1, \ldots, X_n$ in terms of $Y_1, \ldots, Y_n$ as $X_i = s_i(Y_1, \ldots, Y_n)$, then the joint pdf for $Y_1, \ldots, Y_n$ is

$$f[s_1(y_1, \ldots, y_n), \ldots, s_n(y_1, \ldots, y_n)]\, |J(y_1, \ldots, y_n)|$$

where $J(y_1, \ldots, y_n)$ is the n by n determinant with the quantity

$$\frac{\partial s_i(y_1, \ldots, y_n)}{\partial y_j}$$

in the ith row and jth column.

If $f(x_1, \ldots, x_n)$ is the joint pdf for $X_1, \ldots, X_n$, then the joint marginal pdf for X_1, X_2, X_3, for example, is equal to

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4, \ldots, x_n)\, dx_4 \cdots dx_n$$

and is denoted by $f_{1,2,3}(x_1,x_2,x_3)$. The marginal pdf for X_1 is

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_2 \cdots dx_n$$

$X_1, X_2, \ldots, X_n$ are independent if and only if $f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n)$ for all values of $x_1, x_2, \ldots, x_n$. If $X_1, X_2, \ldots, X_n$ are independent and $g_1(x_1), \ldots, g_n(x_n)$ are given functions, it is easily shown that

$$E\{g_1(X_1)g_2(X_2) \cdots g_n(X_n)\} = E\{g_1(X_1)\} E\{g_2(X_2)\} \cdots E\{g_n(X_n)\}$$

If $f(x_1, \ldots, x_n)$ is the joint pdf for $X_1, \ldots, X_n$, then the "joint conditional pdf for X_1, X_2, X_3, given that $X_4 = x_4$ and $\cdots$ and $X_n = x_n$" is defined as

$$\frac{f(x_1, \ldots, x_n)}{f_{4,5,\ldots,n}(x_4, x_5, \ldots, x_n)}$$

and is denoted by $f_{1,2,3}(x_1,x_2,x_3 \mid X_4 = x_4, \ldots, X_n = x_n)$. If $X_1, \ldots, X_n$ are independent, then $f_{1,2,3}(x_1,x_2,x_3 \mid X_4 = x_4, \ldots, X_n = x_n) = f_{1,2,3}(x_1,x_2,x_3)$ for all values of $x_1, \ldots, x_n$.

We often encounter the following problem. $X_1, \ldots, X_n$ have the given joint pdf $f(x_1, \ldots, x_n)$. $Y_1, \ldots, Y_m$ are m chance variables defined as given functions of $X_1, \ldots, X_n$, where $m < n$. The problem is to find the pdf for $Y_1, \ldots, Y_m$. If m were equal to n, we could use the methods already given to solve this problem. Suppose $Y_i = r_i(X_1, \ldots, X_n)$ for $i = 1, \ldots, m$. To solve the problem, we introduce $n - m$ conveniently chosen extra chance variables $Z_1, \ldots, Z_{n-m}$ by the relations $Z_i = s_i(X_1, \ldots, X_n)$ for $i = 1, \ldots, n - m$, where the functions $s_i(x_1, \ldots, x_n)$ are some conveniently chosen ones. Then, by methods already described, we can find the joint pdf for $Y_1, \ldots, Y_m, Z_1, \ldots, Z_{n-m}$, say, $g(y_1, \ldots, y_m, z_1, \ldots, z_{n-m})$. Finally, the required joint pdf for $Y_1, \ldots, Y_m$ is given by

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, \ldots, y_m, z_1, \ldots, z_{n-m})\, dz_1 \cdots dz_{n-m}$$

As an example, suppose $X_1, \ldots, X_n$ are independent, each having the marginal pdf $f(x) = e^{-x}$ for $x > 0$, $f(x) = 0$ for $x \leqslant 0$. We are to find the pdf for $Y = X_1 + \cdots + X_n$. The joint pdf for $X_1, \ldots, X_n$ is $e^{-(x_1 + \cdots + x_n)}$ if $x_1 > 0$ and $\cdots$ and $x_n > 0$ and is equal to zero otherwise. We define the chance variables $Z_1, \ldots, Z_{n-1}$ by $Z_1 = X_1$, $Z_2 = X_1 + X_2, \ldots, Z_{n-1} = X_1 + \cdots + X_{n-1}$. Then the joint pdf for $Z_1, \ldots, Z_{n-1}, Y$ is found to be e^{-y} for $0 < z_1 < z_2 < \cdots < z_{n-1} < y$, and zero otherwise. Then the pdf for Y is zero if $y \leqslant 0$, and for $y > 0$ is given by

$$\int_0^y \int_0^{z_{n-1}} \cdots \int_0^{z_3} \int_0^{z_2} e^{-y}\, dz_1\, dz_2 \cdots dz_{n-2}\, dz_{n-1} = \frac{y^{n-1}e^{-y}}{(n-1)!}$$

Chapter 4

SOME IMPORTANT EXPECTED VALUES AND DISTRIBUTIONS

4.1. Moments and the Moment Generating Function. If X is a chance variable and r and c are two given constants, then $E\{(X-c)^r\}$ is called "the rth moment of X about c," if this expected value exists. The rth moment of X about zero is usually just called "the rth moment." The first moment of X, $E\{X\}$, is also called "the mean of X," and the second moment of X about $E\{X\}$ is called "the variance of X." The variance of X is

$$\begin{aligned} E\{[X-E\{X\}]^2\} &= E\{X^2 - 2E\{X\}X + [E\{X\}]^2\} \\ &= E\{X^2\} - 2[E\{X\}]^2 + [E\{X\}]^2 = E\{X^2\} - [E\{X\}]^2 \end{aligned}$$

The positive square root of the variance of X is called "the standard deviation of X."

If X is a chance variable and t is a given value, then $E\{e^{tX}\}$, as a function of t, is called "the moment generating function for X" and will usually be denoted by $M_X(t)$. It is important to realize that $M_X(t)$ is a function of the ordinary variable t, and not of the chance variable X, though the form of the function is determined by the distribution of X.

If $E\{X^s\}$ exists for every positive integer s, then $E\{X^s\} = d^sM_X(t)/dt^s]_{t=0}$. (This is the reason for the name "moment generating function.") First we prove this when X has only a finite number of possible values, with probability distribution given in general terms as in Sec. 2.2:

Possible values	x_1	x_2	$\cdots$	x_k
Probabilities	p_1	p_2	$\cdots$	p_k

(In this case, $E\{X^s\}$ exists for all positive values of s.) We have

$$M_X(t) = E\{e^{tX}\} = \sum_{i=1}^{k} p_i e^{tx_i}$$

and it is easily seen that

$$d^sM_X(t)/dt^s = \sum_{i=1}^{k} p_i x_i^s e^{tx_i}$$

so that

$$d^sM_X(t)/dt^s]_{t=0} = \sum_{i=1}^{k} p_i x_i^s = E\{X^s\}$$

which completes the proof when X has only a finite number of possible values. When X has an infinite number of possible values which can be listed, as in Sec. 3.1, the demonstration goes through the same way, and the differentiation of the infinite series term by term can be justified by the assumed existence of the moments. In the case where X has a pdf $f(x)$, then

$$M_X(t) = E\{e^{tX}\} = \int_{-\infty}^{\infty} f(x)e^{tx}\,dx$$

and

$$d^sM_X(t)/dt^s = \int_{-\infty}^{\infty} x^s f(x)e^{tx}\,dx$$

so that

$$d^sM_X(t)/dt^s]_{t=0} = \int_{-\infty}^{\infty} x^s f(x)\,dx = E\{X^s\}$$

(Differentiating under the integral sign can be justified by the assumed existence of the moments.)

Although it is sometimes very convenient to compute moments by differentiating moment generating functions, a more important use of the moment generating function is that the distribution of X can often be discovered by recognition of the moment generating function for X. We shall discuss many such cases and need the following two theorems, which we state without proof:

Theorem 1. *If $M(t)$ is a moment generating function for some chance variable and $M(t)$ is finite for all values t in some interval $(-h,h)$, where h is some positive value, then there is exactly one probability distribution corresponding to $M(t)$.*

Theorem 2. *Suppose $M_1(t), M_2(t), \ldots$ is a sequence of moment generating functions, with corresponding cdf's $F_1(x), F_2(x), \ldots$. $M_i(t)$ is assumed finite for all t in some interval $(-h,h)$ for a positive h, for all i. Suppose $M(t)$ is a moment generating function with corresponding cdf $F(x)$, and suppose*

$$\lim_{i\to\infty} M_i(t) = M(t)$$

for each t in the interval $(-h,h)$. Then

$$\lim_{i\to\infty} F_i(x) = F(x)$$

at each x at which $F(x)$ is continuous.

4.2. The Binomial Distribution. If a chance variable X has the possible values 0, 1, 2, . . . , n, where n is a given positive integer, and if $P(X = x) = n!/[x!\,(n - x)!]\,p^x(1 - p)^{n-x}$ for each integer x between 0 and n, where p is a value between 0 and 1, then X is said to have a binomial distribution with parameters n, p. From the discussion in Sec. 1.8, it can be seen that if n independent trials are made, on each of which the outcome E has the probability p of occurring, and if the chance variable X is the number of trials on which E will occur, then X has a binomial distribution with parameters n, p.

If X has a binomial distribution with parameters n, p, then the moment generating function for X, $M_X(t)$ is equal to

$$\sum_{x=0}^{n} \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x}e^{tx} = \sum_{x=0}^{n} \frac{n!}{x!\,(n-x)!}\,(pe^t)^x(1-p)^{n-x}$$
$$= (pe^t + 1 - p)^n$$

Differentiating this last expression, we find $E\{X\} = dM_X(t)/dt]_{t=0} = np$, and $E\{X^2\} = d^2M_X(t)/dt^2]_{t=0} = n(n-1)p^2 + np$, so that the variance of X is equal to $n(n-1)p^2 + np - (np)^2 = np(1-p)$.

4.3. The Poisson Distribution. If a chance variable X has possible values 0, 1, 2, . . . , and $P(X = x) = \lambda^x e^{-\lambda}/x!$ for each nonnegative integer x, λ being a positive value, then X is said to have a Poisson distribution with parameter λ.

$$M_X(t) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}\,e^{tx} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} = e^{-\lambda}\exp(\lambda e^t)$$

from which $E\{X\} = \lambda$, $E\{X^2\} = \lambda + \lambda^2$, and the variance of X is equal to λ.

If X has a binomial distribution with parameters n, p, and if n is large and p is small, then the distribution of X is "almost" a Poisson distribution with parameter equal to np. To see this, suppose X has a binomial distribution with parameters n, p, assume n is large and p is small, and denote np by m. Choose a positive integer x. Then

$$P(X = x) = \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x}$$
$$= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{m}{n}\right)^x\left(1-\frac{m}{n}\right)^{n-x}$$
$$= \frac{1}{x!}\left[\frac{n}{n}\,\frac{n-1}{n}\cdots\frac{n-x+1}{n}\right]m^x\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-x}$$

But if n is large, each factor in the square bracket is close to 1, $(1 - m/n)^{-x}$ is close to 1, and $(1 - m/n)^n$ is close to e^{-m}. Therefore $P(X = x)$ is close to $m^x e^{-m}/x!$, which is the probability for a Poisson distribution with parameter m.

The result in the preceding paragraph can be stated more exactly as follows: if n increases and p decreases with $np = m$ (a fixed positive value) then $\{n!/[x!\,(n - x)!]\}\,p^x(1 - p)^{n-x}$ approaches $m^x e^{-m}/x!$. And the fact that the binomial distribution approaches the Poisson distribution as n increases with $np = m$ (m fixed) also follows from Theorem 2 of Sec. 4.1. For the moment generating function for the binomial distribution is $(1 + pe^t - p)^n = [1 + (m/n)(e^t - 1)]^n$, and this last expression approaches $\exp[m(e^t - 1)]$ as n increases, $\exp[m(e^t - 1)]$ being the moment generating function for the Poisson distribution with parameter m.

Chance variables with Poisson distributions are often encountered in practice. For example, it has been found that the number of persons who will contract a rare noncontagious disease in a given time period has a Poisson distribution, at least approximately. This is often explained as follows. Each person is to be considered a "trial," and if the person contracts the disease in the given time period, we say the outcome E occurs on the trial. Assume each person has the same probability (p, say) of contracting the disease in the time period. Since the disease is noncontagious, the trials represented by different people are independent. If n denotes the total number of people under consideration and X denotes the number of people who will contract the disease in the given time period, then X has a binomial distribution with parameters n, p. But n is large and p is small (it is a *rare* disease), and so X has a Poisson distribution, at least approximately.

4.4. The Hypergeometric Distribution. Suppose we have a box containing R red chips and B black chips, and we mix the chips thoroughly and take out n chips, where $n \leqslant R + B$. Let X denote the number of red chips that will be found among the n chips chosen. We want to find the probability distribution for X. Clearly, the possible values of X are the integers between max $(0, n - B)$ and min (R,n). If x is any integer in this range, then

$$P(X = x) = \frac{\dfrac{R!}{x!\,(R - x)!}\,\dfrac{B!}{(n - x)!\,(B - n + x)!}}{\dfrac{(R + B)!}{n!\,(R + B - n)!}}$$

To see this, let us imagine that each chip is distinguished from all the others in some convenient way (by labeling each with its own number, perhaps). Any particular set of n chips has the same probability of being drawn as any other particular set of n chips, since we assumed thorough mixing. Therefore the probability of getting exactly x red chips is the proportion of sets of n chips containing exactly x red chips. The total number of different sets of n chips is $(R + B)!/[n!\,(R + B - n)!]$, while the number of sets of n chips containing exactly x red chips is the number of ways of picking x red chips out of R multiplied by the number of ways of picking $n - x$ black chips out of B, the product being equal to

$$\frac{R!}{x!\,(R - x)!}\,\frac{B!}{(n - x)!\,(B - n + x)!}$$

If a chance variable X has the probability distribution just described, it is said to have a hypergeometric distribution with parameters n, R, B.

A hypergeometric distribution with n small compared with $R + B$ is "almost" a binomial distribution with parameters n, $R/(R + B)$. To see this intuitively, think of drawing the n chips from the box one by one. Since n is small compared with the total number $R + B$ of chips, the composition of the box will not change very much, and at each drawing the probability of getting a red chip will be approximately $R/(R + B)$. But then the number of red chips that will be drawn has almost the same distribution as the number of heads that will appear in n tosses of a coin with probability of a head on each toss equal to $R/(R + B)$. This latter distribution is binomial with parameters n, $R/(R + B)$.

4.5. The Uniform Distribution. If a chance variable X has a pdf $f(x)$ given as follows:

$$f(x) = \frac{1}{B - A} \qquad \text{for } A < x < B$$

$$f(x) = 0 \qquad \text{for } x < A \text{ or } x > B$$

then X is said to have a uniform distribution between A and B.

Suppose that Y is a chance variable with a continuous cdf $G(y)$ and we define the chance variable Z as $G(Y)$. Then Z has a uniform distribution between 0 and 1. To see this, first we note that Z can take values between 0 and 1 and no others. Fix a value z between 0 and 1, and define $y(z)$ as the largest value y for which $G(y) = z$; that is, $G[y(z)] = z$, but $G[y(z) + \Delta] > z$ for every positive value of Δ. Then the event $(Z \le z)$ is the same event as $[G(Y) \le z]$, which in turn is the same event as $[Y \le y(z)]$. Therefore $P(Z \le z) = P[Y \le y(z)] = G[y(z)] = z$. Thus $K(z)$, the cdf for Z, is given as follows:

$$K(z) = 0 \qquad \text{for } z < 0$$

$$K(z) = z \qquad \text{for } 0 < z < 1$$

$$K(z) = 1 \qquad \text{for } z > 1$$

Differentiating, we find that $k(z)$, the pdf for Z, is

$$k(z) = 1 \qquad \text{for } 0 < z < 1$$

$$k(z) = 0 \qquad \text{for } z < 0 \text{ or } z > 1$$

that is, Z has a uniform distribution between 0 and 1.

4.6. The Normal Distribution. If a chance variable X has the pdf $f(x)$ given as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

where σ is positive, then X is said to have a normal distribution with parameters u, σ. If X has a normal distribution with $u = 0$ and $\sigma = 1$, X is said to have a standard normal distribution.

If X has a normal distribution with parameters u, σ, then

$$M_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2 + tx}\, dx$$

To evaluate this integral, we make the change of variable $y = (x - u)/\sigma$, getting

$$M_X(t) = e^{ut+(1/2)\sigma^2t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1/2)(y-\sigma t)^2}\, dy$$

Evaluating the integral in the last expression by making the change of variable $z = y - \sigma t$, we find $M_X(t) = e^{ut+(1/2)\sigma^2t^2}$. Differentiating, we find that $E\{X\} = u$, $E\{X^2\} = \sigma^2 + u^2$, so that the variance of X is σ^2.

If X has a normal distribution with mean u and variance σ^2 and A and B are constants, then the chance variable Y defined as $AX + B$ has a normal distribution with mean $Au + B$ and variance $A^2\sigma^2$. To show this, we compute the moment generating function for Y, $M_Y(t)$ as follows:

$$\begin{aligned} E\{e^{tY}\} &= E\{e^{t(AX+B)}\} = E\{e^{tB}e^{AtX}\} \\ &= e^{tB}E\{e^{AtX}\} = e^{tB}M_X(At) \\ &= e^{tB}e^{uAt+(1/2)A^2t^2\sigma^2} \\ &= e^{(Au+B)t+(1/2)(A^2\sigma^2)t^2} \end{aligned}$$

and this last expression is seen to be the moment generating function for a chance variable with a normal distribution with mean $Au + B$ and variance $A^2\sigma^2$.

If X has a normal distribution, the value of $P(X \leq x)$ may be found from Table I in the Appendix. Thus, suppose X has a normal distribution with mean u and variance σ^2, and we want to find $P(X \leq c)$, which is given by

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}\, dx$$

To evaluate this integral, we make the change of variable $y = (x - u)/\sigma$, and find that

$$P(X < c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(c-u)/\sigma} e^{-(y^2/2)}\, dy$$

But Table 1 in the Appendix gives values of $(1/\sqrt{2\pi})\int_{-\infty}^{z} e^{-(y^2/2)}\, dy$, for various values of z. As a numerical example, if $u = 2$, $\sigma = 5$, and $c = 10$, we enter Table 1 in the Appendix with $z = (c - u)/\sigma = 1.6$, and find $P(X < c) = 0.945$.

4.7. The Central Limit Theorem. It has long been customary to assume that many chance variables encountered in nature have probability distributions which are approximately normal. In this section we try to explain the basis for this assumption.

First we develop two simple facts about moment generating functions, in the following two lemmas.

Lemma 1. If the chance variable Z has moment generating function $M_Z(t)$, and if A and B are constants, then the moment generating function of the chance variable $AZ + B$ is equal to $e^{tB}M_Z(At)$.

Proof. See the third paragraph of Sec. 4.6.

Lemma 2. If $Z_1, Z_2, \ldots, Z_n$ are independent chance variables, with moment generating functions $M_1(t)$, $M_2(t), \ldots, M_n(t)$, respectively, then the moment generating function for the chance variable $Z_1 + Z_2 + \cdots + Z_n$ is $M_1(t)M_2(t) \cdots M_n(t)$.

Proof. $E\{e^{t(Z_1+Z_2+\cdots+Z_n)}\} = E\{e^{tZ_1}e^{tZ_2}\cdots e^{tZ_n}\} = E\{e^{tZ_1}\}$
$\times E\{e^{tZ_2}\} \cdots E\{e^{tZ_n}\} = M_1(t)M_2(t)\cdots M_n(t)$

If the chance variable Z has moment generating function $M_Z(t)$ and $M_Z(t)$ is finite for all values of t in some interval $(-h,h)$, where h is a positive quantity, then $M_Z(t)$ can be expanded in a Maclaurin expansion, giving

$$M_Z(t) = M_Z(0) + t\, dM_Z(t)/dt]_{t=0} + \frac{t^2}{2!}\, d^2M_Z(t)/dt^2]_{t=0} + \cdots$$

$$= 1 + t\,E\{Z\} + \frac{t^2}{2}\,E\{Z^2\} + \frac{t^2}{2}\Delta(t)$$

where $\Delta(t)$ is a function of t which approaches zero as t approaches zero. We note that if $E\{Z\} = 0$, then the variance of Z is equal to $E\{Z^2\}$.

For the remainder of this section, we assume that $Z_1, Z_2, \ldots$ are independent chance variables, with variances denoted by $\sigma_1^2, \sigma_2^2, \ldots$ and that there are two finite positive numbers A, B, with $A < \sigma_i^2 < B$ for all i. We also assume that the moment generating function for Z_i, denoted by $M_i(t)$, may be written as $1 + t\,E\{Z_i\} + (t^2/2)\,E\{Z_i^2\} + (t^2/2)\,\Delta_i(t)$, and there exists a function $\Delta(t)$ which approaches zero as t approaches zero, such that $|\Delta_i(t)| \leqslant \Delta(t)$ for all values of t and all i. With these assumptions we have the following theorem, known as the "central limit theorem."

Theorem. *If $E\{Z_i\} = 0$ for all i, then as n increases,*

$$P\left(\frac{Z_1 + Z_2 + \cdots + Z_n}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}} < y\right) \quad \textit{approaches} \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-(x^2/2)}\, dx$$

for each value of y.

Proof. For simplicity, we give the proof for the case where $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma^2$, say, and $\Delta_1(t) = \Delta_2(t) = \cdots = \Delta(t)$. The proof in the more general case is practically the same, but the typography becomes complicated. Denote the moment generating function for

$$\frac{(Z_1 + Z_2 + \cdots + Z_n)}{\sqrt{n\sigma^2}}$$

by $\overline{M}_n(t)$. Using Lemmas 1 and 2, we can write

$$\begin{aligned}\overline{M}_n(t) &= M_1\left(\frac{t}{\sigma\sqrt{n}}\right) M_2\left(\frac{t}{\sigma\sqrt{n}}\right) \cdots M_n\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \left[1 + \frac{t^2}{2}\frac{\sigma^2}{n\sigma^2} + \frac{t^2}{2n\sigma^2}\Delta\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n\end{aligned}$$

and as n increases, $\overline{M}_n(t)$ approaches $e^{t^2/2}$ for each t. But $e^{t^2/2}$ is the moment generating function for a standard normal distribution, and Theorem 2 of Sec. 4.1 tells us that the cdf for

$$\frac{(Z_1 + Z_2 + \cdots + Z_n)}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}}$$

approaches the cdf for the standard normal distribution, which completes the proof of the central limit theorem.

Actually, the central limit theorem holds under weaker restrictions than we imposed, but its proof becomes more complicated. However, one restriction that is easily removed is that $E\{Z_i\} = 0$, since the same proof as above serves to show that the cdf of the chance variable

$$\frac{Z_1 + \cdots + Z_n - E\{Z_1\} - \cdots - E\{Z_n\}}{\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}}$$

approaches the standard normal cdf as n increases (noting that the chance variable $Z_i - E\{Z_i\}$ has its mean equal to zero).

One important application of the central limit theorem is to show that under certain conditions, a binomial cdf is close to a normal cdf. If X has a binomial distribution with parameters n, p, then X has the same distribution as the chance variable $Z_1 + Z_2 + \cdots + Z_n$, where $Z_1, Z_2, \ldots, Z_n$ are independent chance variables, each having the following probability distribution:

Possible values	0	1
Probabilities	$1-p$	p

To see this, recall that the distribution of X is the distribution of the number of heads that will be observed in n tosses of a coin which has probability p of coming up head on each individual toss: we can define Z_i as equal to 1 if a head comes up on the ith toss, and equal to 0 otherwise. Then X is equal to $Z_1 + \cdots + Z_n$, and Z_i has the probability distribution tabled above. $E\{Z_i\} = p$, $E\{Z_i^2\} = p$, so the variance of Z_i is $p - p^2$. The central limit theorem states that the cdf for the chance variable $(Z_1 + \cdots + Z_n - np)/\sqrt{n(p - p^2)}$ approaches the standard normal cdf as n increases. But this means that if n is large, the cdf for the chance variable $(X - np)/\sqrt{n(p - p^2)}$ is close to the standard normal cdf, so that $P[(X - np)/\sqrt{n(p - p^2)} < y]$ is closely approximated by $(1/\sqrt{2\pi})\int_{-\infty}^{y} e^{-r^2/2}\, dr$.

On the strength of the central limit theorem, chance variables encountered in nature are often assumed to have a normal distribution because it is felt that they are "built up" as the sum of many "independent components." However, verifying that the exact restrictions of the central limit theorem are satisfied is usually difficult, and a tendency to assume that most chance variables are normally distributed should be resisted.

4.8. The Chi-square Distribution. Before discussing the distribution, we introduce the following standard notation. If k is any positive number, $\Gamma(k)$ is defined as $\int_0^\infty x^{k-1}e^{-x}\, dx$. Integrating by parts, we find that if $k > 1$, $\Gamma(k) = (k-1)\Gamma(k-1)$. Since it is easily verified that $\Gamma(1) = 1$, we have that $\Gamma(k) = (k-1)!$ whenever k is a positive integer. We note that $\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$.

If the chance variable X has the pdf $f(x)$ given as follows:

$$f(x) = \frac{2^{-(k/2)}}{\Gamma(k/2)} x^{k/2-1} e^{-(x/2)} \qquad \text{for } x > 0$$

$$f(x) = 0 \qquad \text{for } x < 0$$

where k is a positive integer, then X is said to have a chi-square distribution with parameter k, or more commonly, X is said to have a chi-square distribution with k degrees of freedom. Thus the term "degrees of freedom" refers to the parameter of a chi-square distribution.

If X has a chi-square distribution with k degrees of freedom;

$$M_X(t) = E\{e^{tX}\} = \frac{2^{-(k/2)}}{\Gamma(k/2)} \int_0^\infty e^{tx} x^{k/2-1} e^{-(x/2)}\, dx$$
$$= \frac{2^{-(k/2)}}{\Gamma(k/2)} \int_0^\infty x^{k/2-1} e^{-(x/2)(1-2t)}\, dx$$

This integral does not exist unless $1 - 2t$ is positive, so for the remainder of this section we assume that $1 - 2t > 0$, or $t < \frac{1}{2}$. Then, making the change of variable $y = (x/2)(1 - 2t)$ in the integral, we find

$$M_X(t) = (1 - 2t)^{-(k/2)} \left[\frac{1}{\Gamma(k/2)} \int_0^\infty y^{k/2-1} e^{-y}\, dy \right] = (1 - 2t)^{-(k/2)}$$

so that $M_X(t) = (1 - 2t)^{-(k/2)}$ if $t < \frac{1}{2}$, and $M_X(t)$ does not exist if $t > \frac{1}{2}$.

Suppose $X_1, X_2, \ldots, X_n$ are independent chance variables, each with a chi-square distribution with $k_1, k_2, \ldots, k_n$ degrees of freedom, respectively. Then the chance variable $Z = X_1 + X_2 + \cdots + X_n$ has a chi-square distribution with $k_1 + k_2 + \cdots + k_n$ degrees of freedom. This is shown easily by finding $M_Z(t)$. By Lemma 2 of Sec. 4.7,

$$M_Z(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t) = (1 - 2t)^{-k_1/2} (1 - 2t)^{-k_2/2} \cdots (1 - 2t)^{-k_n/2}$$
$$= (1 - 2t)^{-(\frac{1}{2})(k_1 + k_2 + \cdots + k_n)}$$

if $t < \frac{1}{2}$. But this is the moment generating function for a chance variable with a chi-square distribution with $k_1 + k_2 + \cdots + k_n$ degrees of freedom, so that by Theorem 1 of Sec. 4.1, Z has a chi-square distribution with $k_1 + k_2 + \cdots + k_n$ degrees of freedom.

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent chance variables, each with a standard normal distribution. Then the chance variable $W = Y_1^2 + \cdots + Y_n^2$ has a chi-square distribution with n degrees of freedom. To show this, we note that by the preceding paragraph it suffices to show that the distribution of Y_i^2 is a chi-square distribution with 1 degree of freedom. The pdf for Y_i is $(1/\sqrt{2\pi})e^{-(\frac{1}{2})y^2}$, and therefore by the discussion in Sec. 3.8, the pdf for $R = Y_i^2$ is

$$\frac{1}{\sqrt{2\pi}} r^{-\frac{1}{2}} e^{-r/2} = \frac{2^{-\frac{1}{2}}}{\Gamma(\frac{1}{2})} r^{\frac{1}{2}-1} e^{-r/2} \qquad \text{for } r > 0$$
$$= 0 \qquad \text{for } r < 0$$

Therefore R has a chi-square distribution with 1 degree of freedom, and W has a chi-square distribution with n degrees of freedom. $E\{Y_i^2\} = 1$, and the variance of Y_i^2 is 2. If n is large, the central limit theorem tells us that the cdf for the chance variable $(W - n)/\sqrt{2n}$ is close to the standard normal cdf.

If W has a chi-square distribution with n degrees of freedom, Table 2 in the Appendix gives the value of w for which $P(W \leq w) = A$, for selected values of A and n. As a numerical example, if $n = 3$, then the value of w for which $P(W \leq w) = 0.95$ is 7.815.

4.9. The Noncentral Chi-square Distribution. Suppose $X_1, X_2, \ldots, X_n$ are independent chance variables, X_i having a normal distribution with mean m_i and variance equal to 1. Define the chance variable W as $X_1^2 + X_2^2 + \cdots + X_n^2$. From Sec. 4.8, we know that if $m_1 = m_2 = \cdots = m_n = 0$, then W has a chi-square distribution with n degrees of freedom. But we are interested in the distribution of W when each m_i is not necessarily zero.

First we find the moment generating function for X_i^2.

$$E\{e^{tX_i^2}\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx^2} e^{-(1/2)(x-m_i)^2}\,dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\sqrt{1-2t}\,x - \frac{m_i}{\sqrt{1-2t}}\right)^2\right]$$

$$\times \exp\left[\frac{1}{2}\left(\frac{m_i^2}{1-2t} - m_i^2\right)\right]dx$$

or, making the change of variable $y = \sqrt{1-2t}\,x - m_i/\sqrt{1-2t}$, we find

$$E\{e^{tX_i^2}\} = e^{\frac{tm_i^2}{1-2t}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(1-2t)^{-1/2}e^{-y^2/2}\,dy$$

$$= (1-2t)^{-1/2}e^{\frac{tm_i^2}{1-2t}} \qquad \text{if } t < \tfrac{1}{2}$$

Therefore

$$E\{e^{tW}\} = (1-2t)^{-n/2}\exp\left[\frac{t}{1-2t}(m_1^2 + \cdots + m_n^2)\right] \qquad \text{if } t < \tfrac{1}{2}$$

This moment generating function shows that the distribution of W does not depend on the individual values of $m_1, \ldots, m_n$, but only on $m_1^2 + \cdots + m_n^2 = m$, say. If $m = 0$, $E\{e^{tW}\} = (1-2t)^{-n/2}$, the moment generating function for a chi-square distribution with n degrees of freedom. If m is greater than zero, W is said to have a noncentral chi-square distribution with n degrees of freedom and noncentrality parameter m.

It is not possible to express the pdf or cdf for a chance variable with a noncentral chi-square distribution in any simple form. However, it can be shown without difficulty that if W has a noncentral chi-square distribution with parameters n and m, and if w is any finite positive number, $P(W \leqslant w)$ decreases as m increases. To show this, we can assume $m_1 = \sqrt{m}, m_2 = \cdots = m_n = 0$, and note that $P(X_1^2 \leqslant x)$ decreases as m_1 increases, for any positive value x.

4.10. The *t* Distribution. Suppose X, Y are independent chance variables, X having a standard normal distribution, Y having a chi-square distribution with n degrees of freedom. We define the chance variable T as $\sqrt{n}X/\sqrt{Y}$ and want to find the pdf for T. To do this, we follow the method described in Sec. 3.12 and introduce the extra chance variable W, defined as equal to X. Next we find the joint pdf for T, W, which we denote by $g(t,w)$. Since X and Y are independent, the joint pdf for X, Y, say, $f(x,y)$, is the product of the marginal pdf's, or

$$f(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{2^{-(n/2)}}{\Gamma(n/2)} y^{n/2-1} e^{-y/2} \qquad \text{for } y > 0$$

$$f(x,y) = 0 \qquad \text{for } y < 0$$

Since $X = W$ and $Y = nW^2/T^2$, we have that the determinant $J(w,t)$ is

$$\begin{vmatrix} \dfrac{\partial x}{\partial w} & \dfrac{\partial x}{\partial t} \\ \dfrac{\partial y}{\partial w} & \dfrac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \dfrac{2nw}{t^2} & -\dfrac{2nw^2}{t^3} \end{vmatrix} = -\frac{2nw^2}{t^3}$$

We note that T and W must have the same sign, since $WT = \sqrt{n}X^2/\sqrt{Y}$. Therefore, if $t > 0$,

$$g(t,w) = \frac{n^{n/2}}{\sqrt{2\pi}\Gamma(n/2)2^{n/2-1}} \frac{w^n}{t^{n+1}} \exp\left[-\frac{w^2}{2}\left(1 + \frac{n}{t^2}\right)\right] \qquad \text{if } w > 0$$

$$= 0 \qquad \text{if } w < 0$$

Therefore, if $t > 0$, the pdf for T, say, $k(t)$, is equal to

$$\int_0^\infty \frac{n^{n/2}}{\sqrt{2\pi}\Gamma(n/2)2^{n/2-1}} \frac{w^n}{t^{n+1}} \exp\left[-\frac{w^2}{2}\left(1 + \frac{n}{t^2}\right)\right] dw$$

and making the change of variable $z = (w^2/2)(1 + n/t^2)$, we find that

$$k(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \qquad \text{if } t > 0$$

If $t < 0$, then the absolute value of $J(w,t)$ is $-(2nw^2/t^3)$, and $g(t,w)$ is given by

$$g(t,w) = -\frac{n^{n/2}}{\sqrt{2\pi}\Gamma(n/2)2^{n/2-1}} \frac{w^n}{t^{n+1}} \exp\left[-\frac{w^2}{2}\left(1 + \frac{n}{t^2}\right)\right] \qquad \text{if } w < 0$$

$$= 0 \qquad \text{if } w > 0$$

Therefore, if $t < 0$, $k(t)$ is equal to

$$-\int_{-\infty}^{0} \frac{n^{n/2}}{\sqrt{2\pi}\Gamma(n/2)2^{n/2-1}} \frac{w^n}{t^{n+1}} \exp\left[-\frac{w^2}{2}\left(1+\frac{n}{t^2}\right)\right] dw$$

and we find that

$$k(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad \text{if } t < 0$$

A chance variable with the pdf $k(t)$ is said to have a *t distribution with n degrees of freedom.*

If T has a t distribution with n degrees of freedom, Table 3 in the Appendix gives the value of t, for which $P(-t \leq T \leq t) = A$, for various values of n and A. For example, if $n = 5$ and $A = 0.90$, the desired value of t is 2.015.

4.11. The *F* Distribution. Suppose the chance variables X, Y are independent, each having a chi-square distribution, with r degrees of freedom for X, s degrees of freedom for Y. We define the chance variable Z as sX/rY. We introduce the extra chance variable W, defined as equal to Y, and find the joint pdf $g(w,z)$ of W, Z. Since $X = rZW/s$ and $Y = W$, the determinant $J(w,z)$ is

$$\begin{vmatrix} \dfrac{rz}{s} & \dfrac{rw}{s} \\ 1 & 0 \end{vmatrix} = -\frac{rw}{s}$$

Also, X, Y are independent, and so the joint pdf $f(x,y)$ for X, Y is

$$f(x,y) = \frac{2^{-\frac{r+s}{2}}}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)} x^{\frac{r}{2}-1} y^{\frac{s}{2}-1} e^{-(\frac{1}{2})(x+y)} \qquad \text{for } x > 0 \text{ and } y > 0$$

$$= 0 \qquad \text{for } x < 0 \text{ or } y < 0$$

Then

$$g(w,z) = \frac{\left(\frac{r}{s}\right)^{\frac{r}{2}} 2^{-\frac{r+s}{2}}}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)} z^{\frac{r}{2}-1} w^{\frac{r+s}{2}-1} e^{-\frac{w}{2}\left(1+\frac{rz}{s}\right)} \qquad \text{for } w > 0 \text{ and } z > 0$$

$$= 0 \qquad \text{for } w < 0 \text{ or } z < 0$$

The pdf for Z, say, $h(z)$, is, for positive z, equal to

$$\int_0^\infty \frac{\left(\frac{r}{s}\right)^{\frac{r}{2}} 2^{-\frac{r+s}{2}}}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)} z^{\frac{r}{2}-1} w^{\frac{sr+}{2}-1} e^{-\frac{w}{2}\left(1+\frac{rz}{s}\right)} \, dw$$

and making the change of variable $v = (w/2)(1 + rz/s)$,

$$h(z) = \frac{\left(\frac{r}{s}\right)^{\frac{r}{2}} 2^{-\frac{r+s}{2}} z^{\frac{r}{2}-1}}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)} \left(\frac{2}{1+\frac{rz}{s}}\right)^{\frac{r+s}{2}} \int_0^\infty v^{\frac{r+s}{2}-1} e^{-v} \, dv$$

$$= \frac{\Gamma\left(\frac{r+s}{2}\right) r^{\frac{r}{2}} s^{\frac{s}{2}}}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)} \frac{z^{\frac{r}{2}-1}}{(s+rz)^{\frac{r+s}{2}}} \qquad \text{for } z > 0$$

$$= 0 \qquad \text{for } z < 0$$

A chance variable with the pdf $h(z)$ is said to have an F distribution with r degrees of freedom in the numerator and s degrees of freedom in the denominator.

If Z has an F distribution with r degrees of freedom in the numerator and s degrees of freedom in the denominator, Table 4 in the Appendix gives the value of z, for which $P(Z \leq z) = A$, for various values of r, s, and A.

4.12. The Noncentral F Distribution. Suppose X, Y are independent chance variables, X having a noncentral chi-square distribution with r degrees of freedom and noncentrality parameter m, and Y having a chi-square distribution with s degrees of freedom. Then the distribution of the chance variable $Z = sX/rY$ is called a noncentral F distribution with r degrees of freedom in the numerator, s degrees of freedom in the denominator, and noncentrality parameter m. It is impossible to write the cdf or the pdf for Z in any simple form, but it can be shown that if z is any finite positive number, $P(Z \leq z)$ decreases as the parameter m increases.

4.13. Moments in the Multivariate Case. If X, Y, Z, . . . are jointly distributed chance variables, we have various "mixed moments," such as $E\{(X-2)^3(Y-5)^2(Z-7)^{20}\}$, etc. In particular, the moment $E\{(X - E\{X\})(Y - E\{Y\})\}$, if it exists, is called the "covariance between

X and Y" and is usually denoted by σ_{XY}. We note that $\sigma_{XY} = E\{XY - Y\,E\{X\} - X\,E\{Y\} + E\{X\}E\{Y\}\} = E\{XY\} - E\{X\}E\{Y\}$. If $\sigma_{XY} = 0$, X and Y are said to be "uncorrelated." If X and Y are independent, they are uncorrelated; but uncorrelated chance variables are not necessarily independent.

The standard deviation of X is commonly denoted by σ_X, and the standard deviation of Y by σ_Y. Assuming that the quantities involved exist, $\sigma_{XY}/\sigma_X\sigma_Y$ is called "the correlation coefficient between X and Y" and is usually denoted by ρ_{XY}. ρ_{XY} can never be below -1 or above 1, as is shown by the following argument. For any given value u, the chance variable $[u(X - E\{X\}) + (Y - E\{Y\})]^2$ is never negative, and therefore its expected value, which is equal to $u^2\sigma_X{}^2 + 2u\sigma_{XY} + \sigma_Y{}^2$, is nonnegative. But this means that the quadratic equation in u, $u^2\sigma_X{}^2 + 2u\sigma_{XY} + \sigma_Y{}^2 = 0$, has at most one real root, or else $u^2\sigma_X{}^2 + 2u\sigma_{XY} + \sigma_Y{}^2$ would become negative for some values of u. This in turn means that $4\sigma_{XY}{}^2 - 4\sigma_X{}^2\sigma_Y{}^2 \leq 0$, or $\sigma_{XY}{}^2/\sigma_X{}^2\sigma_Y{}^2 \leq 1$, or $\rho_{XY}{}^2 \leq 1$, or $-1 \leq \rho_{XY} \leq 1$.

4.14. Multivariate Moment Generating Functions. If $X_1, X_2, \ldots, X_n$ are jointly distributed chance variables, then $E\{e^{t_1X_1 + t_2X_2 + \cdots + t_nX_n}\}$, as a function of the variables $t_1, t_2, \ldots, t_n$, is called the "joint moment generating function for $X_1, X_2, \ldots, X_n$" and will be denoted by $M_{X_1,X_2,\ldots,X_n}(t_1, t_2, \ldots, t_n)$. If $E\{X_1{}^{r_1}X_2{}^{r_2} \cdots X_n{}^{r_n}\}$ exists for all sets of nonnegative integers $r_1, r_2, \ldots, r_n$, then if $s_1, s_2, \ldots, s_n$ are nonnegative integers,

$$E\{X_1{}^{s_1}X_2{}^{s_2} \cdots X_n{}^{s_n}\} = \frac{\partial^{s_1+s_2+\cdots+s_n}}{\partial t_1{}^{s_1}\,\partial t_2{}^{s_2} \cdots \partial t_n{}^{s_n}} \quad M_{X_1,X_2,\ldots,X_n}$$
$$\times\ (t_1,t_2,\ldots,t_n)]_{t_1=\cdots=t_n=0}$$

Also, we have a generalization of Theorem 1 of Sec. 4.1:

Theorem. If *$M(t_1, \ldots, t_n)$ is a joint moment generating function for some set of n chance variables, and if there is a positive value h such that $M(t_1, \ldots, t_n)$ is finite for all sets of values $(t_1, \ldots, t_n)$ with $-h \leq t_i \leq h$ for $i = 1, \ldots, n$, then there is exactly one probability distribution corresponding to $M(t_1, \ldots, t_n)$.*

If $X_1, X_2, \ldots, X_n$ have the joint moment generating function $M_{X_1,X_2,\ldots,X_n}(t_1,t_2, \ldots, t_n)$, then the joint moment generating function for X_1, X_2 is equal to $M_{X_1,X_2,\ldots,X_n}(t_1,t_2,0, \ldots, 0)$. In general, to find the moment generating function for a subset of $X_1, X_2, \ldots, X_n$, it is only necessary to set the t's corresponding to all other X's equal to zero.

4.15. The Joint Distribution of Certain Functions of Normally Distributed Chance Variables. Suppose $X_1, X_2, \ldots, X_n$ are independent chance variables, each with a normal distribution with mean u and standard deviation σ:

W denotes $\qquad \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$

Z denotes $\qquad \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - W)^2$

We shall show that W and Z are independent chance variables, W having a normal distribution with mean u and standard deviation $\sigma/\sqrt{n}$, Z having a chi-square distribution with $n - 1$ degrees of freedom.

As a first step, we find the joint pdf for the chance variables W, $T_1 = X_1 - W$, $T_2 = X_2 - W, \ldots, T_{n-1} = X_{n-1} - W$. Solving for $X_1, X_2, \ldots, X_n$, we find $X_1 = T_1 + W$, $X_2 = T_2 + W, \ldots, X_{n-1} = T_{n-1} + W, X_n = nW - (T_1 + W) - (T_2 + W) - \cdots - (T_{n-1} + W) = W - (T_1 + T_2 + \cdots + T_{n-1})$. Then the determinant $J(w, t_1, \ldots, t_{n-1})$ is

$$\begin{vmatrix} \frac{\partial x_1}{\partial w} & \frac{\partial x_1}{\partial t_1} & \cdots & \frac{\partial x_1}{\partial t_{n-1}} \\ \frac{\partial x_2}{\partial w} & \frac{\partial x_2}{\partial t_1} & \cdots & \frac{\partial x_2}{\partial t_{n-1}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial x_n}{\partial w} & \frac{\partial x_n}{\partial t_1} & \cdots & \frac{\partial x_n}{\partial t_{n-1}} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & -1 & -1 & \cdots & -1 \end{vmatrix} = C_n$$

where C_n is some value depending only on n, whose exact value does not concern us. The joint pdf for $X_1, X_2, \ldots, X_n$ is the product of their separate pdf's and is therefore

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - u}{\sigma}\right)^2\right]$$

A little algebraic manipulation shows that

$$\sum_{i=1}^{n}\left(\frac{x_i - u}{\sigma}\right)^2 = Z + \left(\frac{W - u}{\sigma/\sqrt{n}}\right)^2$$

and that

$$Z = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n-1} T_i^2 + \left(\sum_{i=1}^{n-1} T_i\right)^2\right]$$

Therefore the joint pdf for $W, T_1, \ldots, T_{n-1}$ is

$$|C_n|\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\left(\frac{w - u}{\sigma/\sqrt{n}}\right)^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n-1} t_i^2 - \left(\sum_{i=1}^{n-1} t_i\right)^2\right]\right\}$$

This joint pdf can be written as the product of two factors, one factor involving only w, the other factor involving only $t_1, \ldots, t_{n-1}$. This shows that the chance variable W is independent of the chance variables $T_1, \ldots, T_{n-1}$, and therefore W must be independent of Z, since Z can be written purely in terms of $T_1, \ldots, T_{n-1}$. The factor in the joint pdf involving w is the marginal pdf for W, and this shows that W has a normal distribution with mean u and standard deviation $\sigma/\sqrt{n}$.

All that remains to be shown is that Z has a chi-square distribution with $n - 1$ degrees of freedom. We show this by finding the moment generating function for Z. Since $X_1, \ldots, X_n$ are independent with normal distributions with parameters u, σ, the chance variables $(X_1 - u)/\sigma, \ldots, (X_n - u)/\sigma$ are independent with standard normal distributions, and

$$\sum_{i=1}^{n} \left(\frac{X_i - u}{\sigma}\right)^2$$

has a chi-square distribution with n degrees of freedom. We have seen that

$$\sum_{i=1}^{n} \left(\frac{X_i - u}{\sigma}\right)^2 = Z + \left(\frac{W - u}{\sigma/\sqrt{n}}\right)^2$$

and that Z and W are independent. Therefore

$$E\left\{\exp\left[t \sum_{i=1}^{n} \left(\frac{X_i - u}{\sigma}\right)^2\right]\right\} = E\{\exp(tZ)\}\, E\left\{\exp\left[t\left(\frac{W - u}{\sigma/\sqrt{n}}\right)^2\right]\right\}$$

or since $[(W - u)/(\sigma/\sqrt{n})]^2$ has a chi-square distribution with 1 degree of freedom, $(1 - 2t)^{-(n/2)} = E\{\exp(tZ)\}(1 - 2t)^{-\frac{1}{2}}$, or $E\{\exp(tZ)\} = (1 - 2t)^{-[(n-1)/2]}$ if $t < \frac{1}{2}$. But $(1 - 2t)^{-[(n-1)/2]}$ is the moment generating function for the chi-square distribution with $n - 1$ degrees of freedom, showing that Z has a chi-square distribution with $n - 1$ degrees of freedom.

From the discussion just completed and from Sec. 4.10, it follows that the chance variable

$$\frac{\sqrt{n - 1}\,\dfrac{W - u}{\sigma/\sqrt{n}}}{\sqrt{Z}} = \frac{\sqrt{n}(W - u)}{\sqrt{\dfrac{\sum_{i=1}^{n}(X_i - W)^2}{n - 1}}}$$

has a t distribution with $n - 1$ degrees of freedom.

4.16. The Bivariate Normal Distribution. If the chance variables X_1, X_2 have the joint pdf

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-u_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-u_1}{\sigma_1}\frac{x_2-u_2}{\sigma_2} + \left(\frac{x_2-u_2}{\sigma_2}\right)^2\right]\right\}$$

where $\sigma_1 > 0$, $\sigma_2 > 0$, $\rho^2 < 1$, then the pair X_1, X_2 is said to have a bivariate normal distribution with parameters u_1, u_2, σ_1, σ_2, ρ. We note that if $\rho = 0$, then X_1 and X_2 are independent, X_1 having a normal distribution with mean u_1 and standard deviation σ_1, and X_2 having a normal distribution with mean u_2 and standard deviation σ_2.

Now we find the joint moment generating function for X_1, X_2.

$$M(t_1,t_2) = E\{e^{t_1X_1+t_2X_2}\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x_1+t_2x_2}f(x_1,x_2)\,dx_1\,dx_2$$

Substituting the function $f(x_1,x_2)$ as given above, and then making the change of variables

$$w = \frac{1}{\sqrt{2}}\frac{x_1-u_1}{\sigma_1} - \frac{1}{\sqrt{2}}\frac{x_2-u_2}{\sigma_2}$$

$$v = \frac{1}{\sqrt{2}}\frac{x_1-u_1}{\sigma_1} + \frac{1}{\sqrt{2}}\frac{x_2-u_2}{\sigma_2}$$

we find

$$M(t_1,t_2) = \frac{e^{t_1u_1+t_2u_2}}{2\pi\sqrt{1-\rho^2}}\left\{\int_{-\infty}^{\infty}\exp\left[\frac{v}{\sqrt{2}}(\sigma_1t_1+\sigma_2t_2) - \frac{v^2}{2(1+\rho)}\right]dv\right\} \times \left\{\int_{-\infty}^{\infty}\exp\left[\frac{w}{\sqrt{2}}(\sigma_1t_1-\sigma_2t_2) - \frac{w^2}{2(1-\rho)}\right]dw\right\}$$

Each of the integrals in this expression can be evaluated by the method used to find the moment generating function in Sec. 4.6. The final result is $M(t_1,t_2) = \exp[u_1t_1 + u_2t_2 + (\frac{1}{2})(\sigma_1^2t_1^2 + \sigma_2^2t_2^2 + 2\rho\sigma_1\sigma_2t_1t_2)]$. From this, we find

$$E\{X_1\} = \frac{\partial}{\partial t_1}M(t_1,t_2)]_{t_1=t_2=0} = u_1$$

$$E\{X_2\} = \frac{\partial}{\partial t_2}M(t_1,t_2)]_{t_1=t_2=0} = u_2$$

$$E\{X_1^2\} = \frac{\partial^2}{\partial t_1^2}M(t_1,t_2)]_{t_1=t_2=0} = \sigma_1^2 + u_1^2$$

$$E\{X_2^2\} = \frac{\partial^2}{\partial t_2^2}M(t_1,t_2)]_{t_1=t_2=0} = \sigma_2^2 + u_2^2$$

$$E\{X_1X_2\} = \frac{\partial^2}{\partial t_1\,\partial t_2}M(t_1,t_2)]_{t_1=t_2=0} = \sigma_1\sigma_2 + u_1u_2$$

Recalling the definitions of variance, covariance, and correlation coefficient, we see that $\sigma_1{}^2$ is the variance of X_1, $\sigma_2{}^2$ is the variance of X_2, and ρ is the correlation coefficient between X_1 and X_2.

Setting $t_2 = 0$ in $M(t_1,t_2)$, we find that the marginal distribution for X_1 is normal with mean u_1 and standard deviation σ_1. Setting $t_1 = 0$ in $M(t_1,t_2)$, we find that the marginal distribution for X_2 is normal with mean u_2 and standard deviation σ_2.

4.17. "Length-of-life" Distributions. Suppose the chance variable X is the total length of life of a piece of equipment about which the following assumption is made: the conditional probability that the equipment fails ("dies") during the time interval $(t, t + \Delta t)$, given that it has not failed before time t, is equal to $r(t)\,\Delta t + q(t,\Delta t)$, where $r(t)$ is a given nonnegative function of t, and $q(t,\Delta t)/\Delta t$ approaches zero as Δt approaches zero, uniformly in t. Time is measured from the moment of "birth" of the equipment. Under this assumption, we are going to find the cdf and pdf for X.

From the description of the problem, it is clear that $P(X < 0) = 0$. Let x be a fixed positive value. We compute $P(X > x)$ as follows. Break the interval $(0,x)$ into $x/\Delta x$ subintervals, each of the subintervals having length Δx. Clearly, $P(X > x) = P$ (no failure in first subinterval and no failure in second subinterval $\cdots$ and no failure in last subinterval). But by our assumption, P (no failure in ith subinterval | no failure earlier) $= 1 - r[(i-1)\,\Delta x]\,\Delta x - q((i-1)\,\Delta x, \Delta x)$. Therefore

$$P(X > x) = \prod_{i=1}^{x/\Delta x} \{1 - r[(i-1)\,\Delta x]\,\Delta x - q((i-1)\,\Delta x, \Delta x)\}$$

Taking logarithms, we get

$$\log P(X > x) = \sum_{i=1}^{x/\Delta x} \log\,\{1 - r[(i-1)\,\Delta x]\,\Delta x - q((i-1)\,\Delta x, \Delta x)\}$$

Expanding each logarithm in a Taylor series, we find

$$\log P(X > x) = -\sum_{i=1}^{x/\Delta x} r[(i-1)\,\Delta x]\,\Delta x + Q(\Delta x)$$

where $Q(\Delta x)$ approaches zero as Δx approaches zero. Letting Δx approach zero, we find

$$\log P(X > x) = -\int_0^x r(x)\,dx$$

or

$$P(X > x) = \exp\left[-\int_0^x r(x)\,dx\right]$$

or

$$P(X \leqslant x) = 1 - \exp\left[-\int_0^x r(x)\,dx\right]$$

This gives the cdf for X. The pdf for X is $r(x) \exp\left[-\int_0^x r(x)\,dx\right]$.

The cdf and pdf that we have found depend upon the function $r(x)$. Looking back at the way $r(x)$ enters the problem, it can be seen that $r(x)$ can be interpreted as the "death rate" at time x. It would seem that in most practical applications, $r(x)$ should be an increasing function of x, at least if x is large. This is true for human length of life. However, for lengths of life of vacuum tubes and electric-light bulbs, it is often assumed that $r(x)$ is a positive constant, say, θ. In this case, the cdf and pdf are $1 - e^{-\theta x}$ and $\theta e^{-\theta x}$, respectively, for $x > 0$. Naturally, the pdf and cdf are both equal to zero for all negative values of x. This distribution is known as "the exponential distribution with parameter θ."

Chapter 5

STATISTICAL DECISION PROBLEMS

5.1. Convex Sets. In this section, we discuss some geometric results that will be useful later.

The representation of any pair of values (x_1,x_2) as a point in a plane with perpendicular axes is familiar. Similarly, any set of n values $(x_1,x_2, \ldots, x_n)$ can be thought of as a "point" in "n-dimensional space."

Given any two points in the plane (y_1,y_2) and (z_1,z_2), any point on the line segment joining (y_1,y_2) and (z_1,z_2) can be written as $(ty_1 + (1-t)z_1, ty_2 + (1-t)z_2)$, where t is some value between 0 and 1; that is, if (w_1,w_2) is any given point on the line segment joining (y_1,y_2) and (z_1,z_2), then it is possible to find a value t between 0 and 1 such that $w_1 = ty_1 + (1-t)z_1$ and $w_2 = ty_2 + (1-t)z_2$. So we may say that the line segment joining (y_1,y_2) and (z_1,z_2) consists of the points given by $(ty_1 + (1-t)z_1, ty_2 + (1-t)z_2)$ as t varies between 0 and 1.

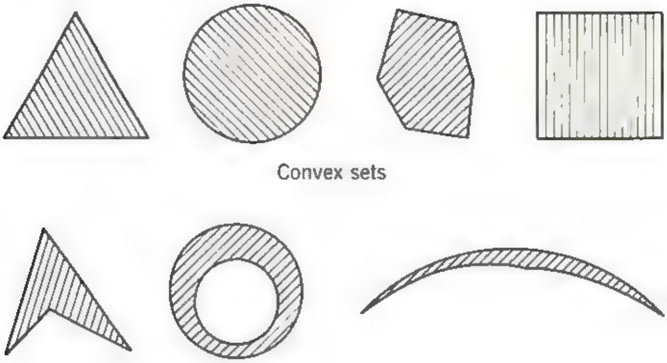
Similarly, if $(y_1,y_2, \ldots, y_n)$ and $(z_1,z_2, \ldots, z_n)$ are any given points in n-dimensional space, then the line segment joining $(y_1,y_2, \ldots, y_n)$ and $(z_1,z_2, \ldots, z_n)$ is defined as the set of points given by $(ty_1 + (1-t)z_1, ty_2 + (1-t)z_2, \ldots, ty_n + (1-t)z_n)$ as t varies between 0 and 1.

If C is a given set of points in n-dimensional space, C is called "convex" if it has the following property: for each and every pair of points in C, all points on the line segment joining the pair of points are in C. Figure 5.1 presents diagrams of convex and nonconvex sets in two-dimensional space.

Suppose $(y_1,y_2, \ldots, y_n)$ and $(z_1,z_2, \ldots, z_n)$ are two given points in n-dimensional space. We say that $(z_1,z_2, \ldots, z_n)$ is "below" $(y_1,y_2, \ldots, y_n)$ if $z_i \le y_i$ for $i = 1, \ldots, n$, with $z_i < y_i$ for at least one value of i. $Q(y_1, \ldots, y_n)$ shall denote the set of all points which are below $(y_1, \ldots, y_n)$. The following theorem (usually called the "supporting hyperplane theorem") is very useful.

Theorem. *If C is a convex set, and $(y_1, y_2, \ldots, y_n)$ is in C but no point in $Q(y_1,y_2, \ldots, y_n)$ is in C, then there are nonnegative numbers $b_1, b_2, \ldots, b_n$, with $b_1 + b_2 + \cdots + b_n = 1$, such that for each and every* *point* $(z_1,z_2,\ldots,z_n)$ *in* C, *we have* $b_1y_1 + b_2y_2 + \cdots + b_ny_n \le b_1z_1 + b_2z_2 + \cdots + b_nz_n$.

Proof. We shall carry out the proof only for the case $n = 2$. The basic idea of the proof is the same when $n > 2$, but the required notation becomes cumbersome. For $n = 2$, we represent the situation in Fig. 5.2.



Convex sets

Nonconvex sets

Fig. 5.1

In Fig. 5.2, we have drawn axes through the point (y_1,y_2), labeled the axes x_1 and x_2, respectively, and labeled the quadrants in the usual way, noting that the quadrant Q_3 is the same as the set of points $Q(y_1,y_2)$.

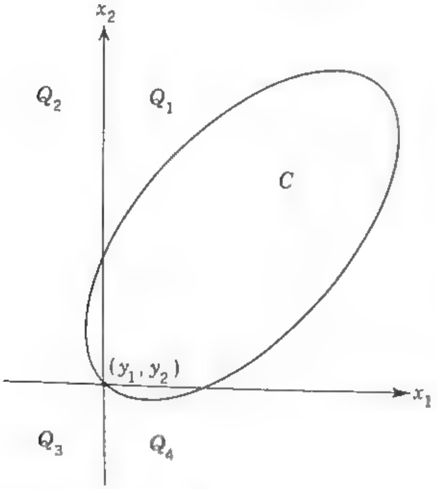


Fig. 5.2

Next we draw a line L_2, starting at (y_1,y_2) and going into Q_2, with the following property: there are no points of C in Q_2 and to the left of L_2, but if L_2 were rotated clockwise however slightly, there would be points of C in Q_2 and to the left of L_2. [If the vertical line through (y_1,y_2) has no points of C to its left, then it is clear that $y_1 \le z_1$ for each and every point (z_1,z_2) in C. This means that the theorem is true with $b_1 = 1$, $b_2 = 0$. Therefore we assume for the remainder of the discussion that L_2 is not vertical.]

Next, we draw a line L_4, starting at (y_1,y_2) and going into Q_4, with the following property: there are no points of C in Q_4 and to the left of L_4, but if L_4 were rotated counterclockwise however slightly, there would be points of C in Q_4 and to the left of L_4. [If the horizontal line through (y_1,y_2) has no points of C below it, then it is clear that $y_2 < z_2$ for each and every point (z_1,z_2) in C. This means that the theorem holds with $b_1 = 0$, $b_2 = 1$. Therefore we assume for the remainder of the discussion that L_4 is not horizontal.]

We now have Fig. 5.3, where θ_3 is the angle between L_2 and L_4. We show that θ_3 cannot be less than 180°. For suppose it were, as in Fig. 5.4. Then, if (p_1,p_2) is any point on L_2 distinct from (y_1,y_2), and (q_1,q_2) is any point on L_4 distinct from (y_1,y_2), the line segment joining (p_1,p_2) and (q_1,q_2) would contain points of $Q(y_1,y_2)$. But there is a point $(\bar{p}_1,\bar{p}_2)$ in C and either on L_2 or arbitrarily close to L_2, and a point $(\bar{q}_1,\bar{q}_2)$ in C and either on L_4 or arbitrarily close to L_4, because of the definition of L_2 and L_4. Then the line segment joining $(\bar{p}_1,\bar{p}_2)$ and $(\bar{q}_1,\bar{q}_2)$ contains points of $Q(y_1,y_2)$, and all the points on this line segment are in C. This contradicts our assumption that no point in $Q(y_1,y_2)$ is in C and proves that θ_3 is at least 180°.

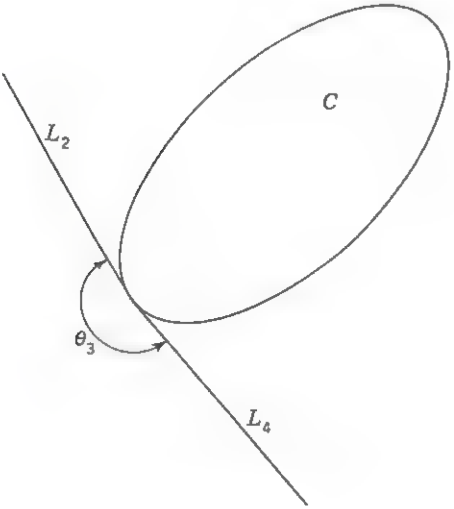


Fig. 5.3

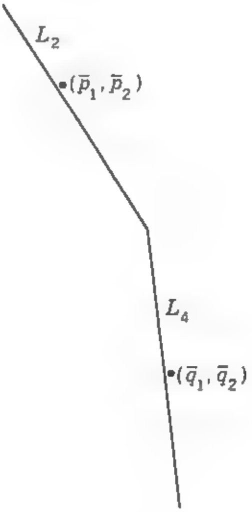


Fig. 5.4

Since θ_3 is at least 180°, we see that if the line L_2 is extended indefinitely in both directions, no points of C are to the left of this extended L_2. The equation of the line L_2 can be written as $a_1x_1 + a_2x_2 + d = 0$, for some constants a_1, a_2, d. Since the slope of L_2 is negative, a_1 and a_2 have the same sign, and $a_1 + a_2$ is not equal to zero. Then we can write the equation of L_2 as

$$\frac{a_1}{a_1 + a_2}x_1 + \frac{a_2}{a_1 + a_2}x_2 = \frac{-d}{a_1 + a_2}$$

Denoting $a_1/(a_1 + a_2)$ by b_1, $a_2/(a_1 + a_2)$ by b_2, $-d/(a_1 + a_2)$ by k, the equation of L_2 can be written $b_1x_1 + b_2x_2 = k$, where b_1, b_2 are nonnegative, and $b_1 + b_2 = 1$. But no points of C are to the left of L_2, and therefore if (z_1,z_2) is any point of C, $b_1z_1 + b_2z_2 \geq k$. Since (y_1,y_2) is on L_2, we have $b_1y_1 + b_2y_2 = k \leq b_1z_1 + b_2z_2$, completing the proof of the theorem for the case $n = 2$.

We note that the values of (b_1,b_2) depend on the particular point (y_1,y_2) we are dealing with. Also, for a fixed point (y_1,y_2), the values of (b_1,b_2) may or may not be uniquely determined: they are uniquely determined if and only if the angle θ_3 is exactly 180°.

As an example, if the set C is a circle and its interior, the line L_2 is simply the tangent to the circle at (y_1,y_2). As another example, if C is a square and its interior, and the point (y_1,y_2) is the lower left-hand corner of the square, L_2 can be the left boundary of the square, or the lower boundary of the square, or any line we get by rotating the lower boundary clockwise around (y_1,y_2) through less than a right angle.

5.2. Description of the General Problem of Statistics. The typical problem that a statistician is called upon to help solve may be briefly described as follows. $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is a set of $m + n$ jointly distributed chance variables. Somebody has to choose one of a given set of possible decisions or actions after having observed $X_1, \ldots, X_m$ but before observing $Y_1, \ldots, Y_n$. It is known that the joint probability distribution of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is one of a given set of possible probability distributions, but exactly which distribution is not known. After a particular decision is chosen, $Y_1, \ldots, Y_n$ will be observed, and a loss will be incurred which depends on the decision chosen and the observed values of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. (A profit is regarded as a negative loss.) The problem is: which decision should be chosen after observing $X_1, \ldots, X_m$?

As an example of such a problem, suppose a company is formed which intends to build a factory to manufacture a new type of perishable foodstuff. The question is what the productive capacity of the factory should be. If the capacity is small, the factory will not cost much to build or operate, but then potential profits will be lost if demand for the product is greater than the capacity of the factory. If the capacity is large, the factory will be expensive to build and operate and may be idle a good part of the time if its capacity exceeds the demand for the product. In order to avoid the pitfalls of over- or undercapacity, it is decided to run a survey to try to measure the potential demand for the product. This survey consists in choosing certain persons at random and finding out how much of the product each would buy. In addition, the trend of population in the area, taxes, the effect of advertising, and other quantities may be observed. Here the possible decisions are the possible capacities of the factory, the chance variables $X_1, \ldots, X_m$ are the various quantities observed, in the survey, and $Y_1, \ldots, Y_n$ are the demands that will be observed after the factory is built; that is, Y_i is the demand that will be observed in the ith accounting period after the factory is built, and n is taken large enough so that the n periods cover the life of the factory. The possible joint distributions of $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ need not concern us at present, although people experienced in running surveys presumably know what sort of joint distributions apply.

5.3. Further Discussion of Statistical Problems. The description of a statistical problem given in Sec. 5.2 differs from the description usually given, in one important respect. In the usual statement of the problem, the loss depends on the decision chosen, on the joint distribution of $X_1, \ldots, X_m$, and on the observed values of $X_1, \ldots, X_m$, but not on $Y_1, \ldots, Y_n$ at all; in fact, the usual problem does not even mention the existence of $Y_1, \ldots, Y_n$. However, it is difficult to think of practical problems in which the usual formulation would be reasonable. For example, let us go back to the problem of deciding what size factory to build. It is clear that what really determines the profitability of the factory is not the distribution of $X_1, \ldots, X_m$, but rather the values of $Y_1, \ldots, Y_n$. Indeed, the only reason we bother observing $X_1, \ldots, X_m$ is to try to learn what the values $Y_1, \ldots, Y_n$ may reasonably be expected to be.

A second reason for preferring the formulation of Sec. 5.2 to the usual formulation is the difficulty of imagining just what mechanism would or could disclose the actual joint distribution of $X_1, \ldots, X_m$ to us, so that we may know what loss we have incurred. Except for certain artificial "game" situations where somebody knew what the actual joint distribution was but deliberately withheld the information until the time arrived for the loss to be incurred, any disclosures about the joint distribution of $X_1, \ldots, X_m$ would be made by means of observing additional chance variables $Y_1, \ldots, Y_n$ which are jointly distributed with $X_1, \ldots, X_m$.

A third reason for preferring the formulation of Sec. 5.2 to the usual formulation is that the usual setup is a special case of the setup of Sec. 5.2, but not vice versa. To see this, note that we could arrange things so that once the values of $Y_1, \ldots, Y_n$ are observed, the joint distribution of $X_1, \ldots, X_m$ is completely known. The artificiality of this procedure simply points up the artificiality of the usual formulation of the problem.

In spite of the objections we have raised to the usual formulation of the statistical problem, most of the published research and discussions use that formulation, and we return to it in Chap. 9.

A point that is worth stressing is that what makes statistical problems particularly difficult is the fact that the joint probability distribution of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is not completely known. Problems where this joint distribution is completely known are not considered statistical problems and are conceptually much easier to handle than statistical problems, though the computations required may be formidable. In many problems where the joint distribution is completely known, it is not necessary to observe $X_1, \ldots, X_m$ before choosing a decision. Problems of this type appear in the exercises of earlier chapters.

5.4. Decision Rules. The statistician's role is usually not that of personally choosing a decision; the choice is a prerogative of the person or persons who are actually going to incur the loss or make the profit. The role of the statistician is to act as an adviser to the decision maker. Suppose a statistician is acting as adviser to a businessman who has to decide whether or not to open a branch office in a certain town. Perhaps the businessman has a strong aversion to the town, for reasons that have nothing to do with the desirability of opening the office there, and the statistician may be completely unaware of this aversion (even the businessman may be largely unaware of it). The effect of the aversion might be to make the businessman decide against opening the office, even in the face of the strongest evidence that opening the office would be desirable. How can the statistician avoid this unfortunate situation? (Of course, the businessman and not the statistician is the one who will suffer from a poor decision, but the statistician may well fear that his reputation would be damaged if he were associated with a disastrous decision.) The standard way of avoiding such a situation is for the statistician to insist that the businessman specify, *before* $X_1, \ldots, X_m$ *are observed*, which observed values would cause him to decide to open the office and which observed values would cause him to decide not to open the office. Presumably, the businessman would recognize the irrationality of specifying that he would not open the office even though the observations showed that it would be very profitable to do so, if he were made to state what he would do before observing the values of $X_1, \ldots, X_m$. However, if he weren't forced to state his intentions before observing the values of $X_1, \ldots, X_m$, he could always claim that the values of $X_1, \ldots, X_m$ actually observed simply weren't encouraging enough to justify opening the office.

Thus a rational analysis of the problem starts with the statistician obtaining from his employer a list showing which decision would be chosen for each possible set of observed values of $X_1, \ldots, X_m$. Such a list is called a "decision rule." The statistician's major task is to judge the goodness of any given decision rule, relative to other possible decision rules.

In the case of the businessman and his problem of whether or not to open an office, suppose $X_1, \ldots, X_m$ are the demands in dollars for the businessman's product of m persons chosen at random in the town. One possible decision rule is to decide to open the office if $\sum_{i=1}^{m} X_i >$ \$10,000; otherwise not to open the office. A different possible decision rule is to decide to open the office if at least one-quarter of the observed values of $X_1, \ldots, X_m$ were above \$25. A great number of other possible decision rules clearly exists.

5.5. Decision Rules Using Randomization. In Sec. 5.4 we defined a decision rule as the assignment of a particular decision to each possible set of observed values of $X_1, \ldots, X_m$. We are now going to generalize this definition to allow a decision rule to be an assignment of a set of probabilities of choosing the various decisions to each possible set of observed values of $X_1, \ldots, X_m$. A decision rule of this more general type for the problem discussed in Sec. 5.4 is: Decide to open the office if $\sum_{i=1}^{m} X_i >$ \$10,000; decide not to open the office if $\sum_{i=1}^{m} X_i <$ \$10,000; if $\sum_{i=1}^{m} X_i =$ \$10,000, assign probability $\frac{1}{4}$ to opening the office and probability $\frac{3}{4}$ to not opening the office. This means that if the sum of the observations equals exactly \$10,000, the businessman chooses his decision by using a random device that assigns probability $\frac{1}{4}$ to opening the office. This decision rule is said to use "randomization."

Decision rules which use randomization strike most people as very peculiar, and hair-raising pictures can be drawn of the fate that would overtake the managers of a corporation if the stockholders found them deciding policy by tossing coins or rolling dice or spinning spinners. We shall see below that randomization is actually used only rarely. The introduction of randomized decision rules serves to simplify the mathematical development.

5.6. Notation for Statistical Decision Problems. A convenient system of notation is a necessity for handling the variety of problems we shall meet, and so we shall develop our notation rather carefully. There are several things to keep track of: the different possible joint distributions of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$; the different possible sets of values of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$; and the different possible decisions.

We shall use the symbol θ as an index for the possible joint distributions; that is, a particular value of θ picks out a particular distribution. In the case where there is only a finite number of possible joint distributions, say, h of them, we can list them in some definite order and let θ range over the integers from 1 to h. Then, when $\theta = i$, it means that we are dealing with the ith distribution in our list. There are more complicated cases. For example, the X's and Y's might be known to be independent chance variables, each with the same Poisson distribution, the parameter of this Poisson distribution being unknown. In this case, θ will correspond to the parameter of the Poisson distribution and will range over all positive numbers. Thus, when $\theta = 2.6$, it means that we are dealing with the Poisson distribution with parameter 2.6

We use the symbol x as an index for the possible sets of values of $X_1, \ldots, X_m$. Thus x ranges over certain points in m-dimensional space. We use the symbol y as an index for the possible sets of values of $Y_1, \ldots, Y_n$. Thus y ranges over certain points in n-dimensional space.

We use the symbol D as an index for the possible decisions; that is, a particular value of D picks out a particular decision. In the case where there is only a finite number of possible decisions, say, L of them, we can list them in some definite order, and let D range over the integers from 1 to L. Then, when $D = i$, it means that we are dealing with the ith decision in our list. There are more complicated cases. For example, we shall come across cases where any value between 0 and ∞ is a possible decision. Then D ranges over all nonnegative numbers.

The loss we incur when the values of $X_1, \ldots, X_m$ are given by x, the values of $Y_1, \ldots, Y_n$ are given by y, and the decision chosen is D will be denoted by $W(y;D;x)$. In many cases, the loss depends only on y and D, and not explicitly on x. In such a case, we write the loss as $W(y;D)$.

When the joint distribution corresponding to θ allows us to list the possible values of the chance variables with their probabilities, $f(x,y;\theta)$ denotes the probability assigned to the $m + n$ dimensional point x, y by this distribution. When the distribution corresponding to θ has a joint pdf, $f(x,y;\theta)$ denotes the value of this joint pdf at the $m + n$ dimensional point x, y.

A decision rule s is defined by nonnegative numbers $s(D;x)$, where $s(D;x)$ is the probability assigned by the decision rule s to choosing the decision D when the point x is observed. Thus, when D can take on only L different values, say, $1, 2, \ldots, L$, we have

$$\sum_{D=1}^{L} s(D;x) = 1 \qquad \text{for each } x$$

For each given decision rule s, the loss that will be incurred when using s is a chance variable whose probability distribution depends upon the unknown joint probability distribution of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. The expected value of the loss that will be incurred when using the decision rule s and when the joint probability distribution is given by θ will be denoted by $r(\theta;s)$. $r(\theta;s)$ is often called "the risk when using s, and the true distribution is given by θ."

If we are dealing with a problem in which there is a finite number L of possible decisions and the joint distribution given by θ allows only a finite number of possible $m + n$ dimensional points x, y, then

$$r(\theta;s) = \sum_{x} \sum_{y} \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)s(D;x)$$

where x and y in the summation run over all possible sets of values of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ allowed by the joint distribution. To see this equality, it is necessary to note only that for a given θ, the possible values of the loss are the values $W(y;D;x)$ takes as y, D, and x vary, and the probability that the value of the loss will be $W(y;D;x)$ is equal to $f(x,y;\theta)\ s(D;x)$, which is the probability that x, y will be observed and that D will be chosen after x is observed. Then the formula for $r(\theta;s)$ follows from the definition of expected value.

If we are dealing with a problem with a finite number L of possible decisions, and the distribution corresponding to θ has a pdf $f(x,y;\theta)$, then

$$r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)s(D;x)\,dx_1 \cdots dx_m\,dy_1 \cdots dy_n$$

where x, y in the integral denote vectors $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, respectively.

We close this section with two simple numerical examples, to illustrate the concepts introduced above.

Example 1. We have to decide whether or not to buy a certain piece of equipment for \$500. The equipment may turn out to be defective or not, and is not guaranteed. The alternative to buying this equipment is to install a different type of guaranteed device for \$1,000. Before deciding, we shall have the opportunity to observe two similar pieces of equipment, produced by the same factory. The proportion of defectives turned out by this factory is unknown. If we buy the equipment and it turns out to be defective, we shall install the guaranteed device at an additional cost of \$1,000.

We introduce the following notation. The chance variable Y is defined to be 1 if the piece of equipment we are considering turns out to be defective, and 0 otherwise. X_1 is defined to be 1 if the first of the two similar pieces of equipment to be observed turns out to be defective, and 0 otherwise. X_2 is defined the same way in terms of the second similar piece of equipment to be observed. θ denotes the unknown proportion of defectives turned out by the factory producing this equipment. Of course, $0 \leq \theta \leq 1$. Then X_1, X_2, Y are independent chance variables, each with the following probability distribution:

Possible values	0	1
Probability	$1 - \theta$	θ

We label the decision to buy as decision number 1 and the decision to use the guaranteed device as decision number 2. The loss function does not depend explicitly on X_1 or X_2, and is seen to be as follows:

$$W(0;1) = 500,\ W(1;1) = 1{,}000 + 500,\ W(0;2) = W(1;2) = 1{,}000$$

Then, using our formula for $r(\theta;s)$ above, we have

$$\begin{aligned} r(\theta;s) = {} & [500(1-\theta)^3 + 1{,}500\theta(1-\theta)^2]s(1;0,0) \\ & + [500\theta(1-\theta)^2 + 1{,}500\theta^2(1-\theta)][s(1;0,1) + s(1;1,0)] \\ & + [500\theta^2(1-\theta) + 1{,}500\theta^3]s(1;1,1) \\ & + 1{,}000[(1-\theta)^3 + \theta(1-\theta)^2]s(2;0,0) \\ & + 1{,}000[\theta(1-\theta)^2 + \theta^2(1-\theta)][s(2;0,1) + s(2;1,0)] \\ & + 1{,}000[\theta^2(1-\theta) + \theta^3]s(2;1,1) \end{aligned}$$

As a particular case, suppose s_1 is the decision rule with $s_1(1;0,0) = 1$, $s_1(1;0,1) = s_1(1;1,0) = \frac{1}{2}$, $s_1(1;1,1) = 0$. Then, recalling that $s_1(2;x) = 1 - s_1(1;x)$ for all x, we find $r(\theta;s_1) = 500 + 1{,}500\theta - 1{,}000\theta^2$. As another case, suppose s_2 is the decision rule with $s_2(1;0,0) = 0$, $s_2(1;0,1) = s_2(1;1,0) = \frac{1}{2}$, $s_2(1;1,1) = 1$. Then we find that $r(\theta;s_2) = 1{,}000 - 500\theta + 1{,}000\theta^2$.

Example 2. A company has to decide on the price it will charge for a certain piece of equipment. It feels that it can charge \$10,000 plus 100 times the length of time it guarantees the equipment (time is measured in months). If it guarantees the equipment for D months, it refunds the purchase price if the equipment fails before D months pass and refunds nothing if the equipment lasts for at least D months. It is felt that the length of life of the equipment has an exponential distribution with unknown parameter θ (Sec. 4.17). Before making its decision, the company will observe the length of life of three similar pieces of equipment.

We introduce the following notation. Y is the length of life of the guaranteed equipment, in months. X_1, X_2, X_3 are the lengths of life of the three similar pieces of equipment that will be observed. X_1, X_2, X_3, Y are independent chance variables, each with pdf equal to $\theta e^{-\theta x}$ for $x > 0$, zero for $x < 0$. D is the length of time the guarantee runs, in months. The loss function does not depend on X_1, X_2, X_3, and is seen to be as follows:

$$\begin{aligned} W(y;D) &= -(10{,}000 + 100D) && \text{if } y \geq D \\ W(y;D) &= 0 && \text{if } y < D \end{aligned}$$

We note that in this problem there is an infinite number of possible decisions, and our formulas for $r(\theta;s)$ given above assume a finite number of possible decisions. However, given a decision rule s, it is a simple matter to compute $r(\theta;s)$. As an example, suppose our decision rule s_1 is to set D equal to $X_1 + X_2 + X_3$. This means that $s_1(D;x_1,x_2,x_3)$ is equal to 1 if $D = x_1 + x_2 + x_3$ and is equal to 0 otherwise. Then we get

$$r(\theta;s_1) = \int_0^\infty \int_0^\infty \int_0^\infty \left[\int_0^{x_1+x_2+x_3} 0\,dy \right.$$
$$\left. + \int_{x_1+x_2+x_3}^\infty \theta^4 e^{-\theta(x_1+x_2+x_3+y)} [-10{,}000 - 100(x_1 + x_2 + x_3)]\,dy \right] dx_1\,dx_2\,dx_3$$

To see this, we go back to our formula for $r(\theta;s)$ and note that since D is a fixed function of x_1, x_2, x_3, the $\sum_D$ operation disappears, and we simply replace D by $x_1 + x_2 + x_3$ in the computation. When the integration is carried out, we find $r(\theta;s_1) = -1{,}250 - 300/16\theta$.

5.7. The Comparison of Decision Rules. It is clear from our discussion so far that in each statistical decision problem there are infinitely many decision rules and that some criterion for characterizing a decision rule as "good" or "bad" is needed.

Roughly speaking, we are going to consider a decision rule s "good" when $r(\theta;s)$ is "small" for all possible values of θ in the problem. Since $r(\theta;s)$ is the expected loss when the decision rule s is used and the distribution corresponding to θ is the actual distribution, this is not an unreasonable criterion for calling a decision rule good. We must consider $r(\theta;s)$ for *all* possible values of θ because we do not know which distribution is the true one.

To be more precise, suppose we are considering two different decision rules, s_1 and s_2, characterized by the decision probabilities $s_1(D;x)$ and $s_2(D;x)$ respectively. Suppose $r(\theta;s_1) \leq r(\theta;s_2)$ for all possible values of θ, with $r(\theta;s_1) < r(\theta;s_2)$ for at least one value of θ. Then we say that s_1 is a better decision rule than s_2, and we would certainly not use the decision rule s_2. A decision rule t is called "inadmissible" if there is a decision rule which is better than t according to the definition of "better" just given. For example, it can be verified that in Example 1 of Sec. 5.6, s_1 is a better decision rule than s_2. This is not surprising, since s_2 specifies that the equipment is bought if both similar pieces of equipment are observed to be defective. Thus s_2 in Example 1 of Sec. 5.6 is inadmissible. Any decision rule which is not inadmissible is called "admissible." It should be emphasized that just because s_1 of Example 1 of Sec. 5.6 has been shown to be better than s_2 does not necessarily mean that s_1 is admissible: there may be a decision rule s_3 which is better than s_1.

Whatever decision rule is finally used should certainly be an admissible decision rule, and therefore it would be useful to develop a method for finding admissible decision rules.

For the remainder of this section, we shall assume that our decision problem contains a finite number h of possible joint probability distributions for $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, so that θ may be assumed to range over the values $1, 2, \ldots, h$. Then any decision rule s has associated with it the h risks $r(1;s), r(2;s), \ldots, r(h;s)$. These risks can be plotted as a point in h-dimensional space. Thus any decision rule s has associated with it a point in h-dimensional space. We denote by C the set of all the points in h-dimensional space associated with some decision rule. We shall show that C is a convex set, by the following argument. Suppose $w_1, \ldots, w_h$ and $z_1, \ldots, z_h$ are any two points in C. Then there are decision rules s and t, such that $r(\theta;s) = w_\theta$ and $r(\theta;t) = z_\theta$ for $\theta = 1, \ldots, h$. Given any value q between 0 and 1, we define a decision rule μ_q whose decision probabilities are $\mu_q(D;x) = qs(D;x) + (1-q)\,t(D;x)$. Using the equation given in Sec. 5.6 for the case of a finite number of possible values for the chance variables $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ (the case where there is a joint pdf is handled analogously), we have

$$\begin{aligned}
r(\theta;\mu_q) &= \sum_x \sum_y \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)\mu_q(D;x) \\
&= \sum_x \sum_y \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)[qs(D;x) + (1-q)t(D;x)] \\
&= q\sum_x \sum_y \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)s(D;x) \\
&\qquad + (1-q)\sum_x \sum_y \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)t(D;x) \\
&= qr(\theta;s) + (1-q)r(\theta;t) = qw_\theta + (1-q)z_\theta \qquad \text{for } \theta = 1, \ldots, h
\end{aligned}$$

Thus the h-dimensional point $qw_1 + (1-q)\,z_1, \ldots, qw_h + (1-q)\,z_h$ is in C, since this point corresponds to the decision rule μ_q. This shows that C is a convex set, since as q varies between 0 and 1, we get the line segment joining the points $w_1, \ldots, w_h$ and $z_1, \ldots, z_h$.

In the terminology of Sec. 5.1, it is clear that if s is an admissible decision rule, no point in C can be below the point $r(1;s), r(2;s), \ldots, r(h;s)$. But then by the theorem of Sec. 5.1, there are h nonnegative numbers $b(1), \ldots, b(h)$, with $b(1) + b(2) + \cdots + b(h) = 1$, such that if t is any decision rule,

$$\begin{aligned}
&b(1)r(1;s) + b(2)r(2;s) + \cdots + b(h)r(h;s) \\
&\qquad\qquad \leqslant b(1)r(1;t) + b(2)r(2;t) + \cdots + b(h)r(h;t)
\end{aligned}$$

We have proved the following theorem:

Theorem. *If s is an admissible decision rule, then there are h nonnegative numbers $b(1), \ldots, b(h)$, with $b(1) + \cdots + b(h) = 1$, such that for each and every decision rule t,*

$$b(1)r(1;s) + \cdots + b(h)r(h;s) \leqslant b(1)r(1;t) + \cdots + b(h)r(h;t)$$

The numbers $b(1), \ldots, b(h)$ depend on the decision rule s and may not be unique.

For a numerical example, we take a case where $h = 2$, so that we can graph the set C. We simplify Example 1 of Sec. 5.6 by assuming that the proportion of defectives turned out by the factory is known to be either ¼ or ½. This requires a change in the interpretation of θ. Now when $\theta = 1$, it will mean that the common probability distribution of X_1, X_2, Y is

Possible values	0	1
Probability	¼	¾

and when $\theta = 2$, it will mean that the common probability distribution is

Possible values	0	1
Probability	½	½

Then, using the equation given in Sec. 5.6, we have

$$\begin{aligned} r(1;s) = {} & (27{,}000/64)s(1;0,0) + (9{,}000/64)[s(1;0,1) + s(1;1,0)] \\ & + (3{,}000/64)s(1;1,1) + (36{,}000/64)[1 - s(1;0,0)] \\ & + (12{,}000/64)[2 - s(1;0,1) - s(1;1,0)] + (4{,}000/64)[1 - s(1;1,1)] \end{aligned}$$

$r(2;s) = 1{,}000$, for any s. In this case the set C consists of the line segment with $r(2;s) = 1{,}000$, $r(1;s)$ ranging from 48,000/64 to 64,000/64. A line segment is a convex set. There is only one admissible decision rule in this problem: it is the decision rule s_1 with $s_1(1;x) = 1$ for all possible x. Then $r(1;s_1) = 48{,}000/64$, $r(2;s_1) = 1{,}000$.

Usually, the convex set C does not degenerate into a line, as it did in the preceding example. As another example, we modify the preceding example by assuming that the proportion of defectives turned out by the factory is known to be either ¼ or ¾. Now when $\theta = 1$, it will mean that the proportion is ¼, and when $\theta = 2$ it will mean that the proportion is ¾. Then using the equation given in Sec. 5.6, we have

$$\begin{aligned} r(1;s) = {} & (27{,}000/64)s(1;0,0) + (9{,}000/64)[s(1;0,1) + s(1;1,0)] \\ & + (3{,}000/64)s(1;1,1) + (36{,}000/64)[1 - s(1;0,0)] \\ & + (12{,}000/64)[2 - s(1;0,1) - s(1;1,0)] + (4{,}000/64)[1 - s(1;1,1)] \\ r(2;s) = {} & (5{,}000/64)s(1;0,0) + (15{,}000/64)[s(1;0,1) + s(1;1,0)] \\ & + (45{,}000/64)s(1;1,1) + (4{,}000/64)[1 - s(1;0,0)] \\ & + (12{,}000/64)[2 - s(1;0,1) - s(1;1,0)] + (36{,}000/64)[1 - s(1;1,1)] \end{aligned}$$

Letting $s(1;0,0)$, $s(1;0,1)$, $s(1;1,0)$, and $s(1;1,1)$ vary independently between 0 and 1 and plotting the two-dimensional points $[r(1;s), r(2;s)]$

resulting, we get Fig. 5.5. The circled points represent decision rules which do not use randomization. {A decision rule s which does not use randomization is one for which $s(D;x)[1 - s(D;x)] = 0$ for all values of D and x.} The following table gives $r(1;s)$ and $r(2;s)$ for all decision rules in our problem not using randomization.

$s(1;0,0)$	$s(1;0,1)$	$s(1;1,0)$	$s(1;1,1)$	$r(1;s)$	$r(2;s)$
1	1	1	1	48,000/64	80,000/64
1	1	0	1	51,000/64	77,000/64
1	0	1	1	51,000/64	77,000/64
1	0	0	1	54,000/64	74,000/64
1	1	1	0	49,000/64	71,000/64
1	1	0	0	52,000/64	68,000/64
1	0	1	0	52,000/64	68,000/64
1	0	0	0	55,000/64	65,000/64
0	1	1	1	57,000/64	79,000/64
0	1	0	1	60,000/64	76,000/64
0	0	1	1	60,000/64	76,000/64
0	0	0	1	63,000/64	73,000/64
0	1	1	0	58,000/64	70,000/64
0	1	0	0	61,000/64	67,000/64
0	0	1	0	61,000/64	67,000/64
0	0	0	0	64,000/64	64,000/64

Note that although 16 decision rules are listed, they represent only 12 distinct points on the diagram. Also note that the boundary of the convex set C consists of line segments joining points corresponding to decision rules not using randomization. We shall prove later that this is always the case, and thus a convenient way to sketch the set C is to plot all the points corresponding to decision rules which do not use randomization, and then draw line segments between pairs of these points in such a way that each of the points is either on one of the line segments or else contained inside the boundary formed by the line segments, and the line segments enclose a convex set. These line segments then form the boundary of the convex set C. The shaded part of the boundary in the diagram represents the admissible decision rules.

As a third example, we simplify Example 2 of Sec. 5.6 by assuming that the unknown parameter is known to be equal to either 0.02 or 0.04 and that D must be one of the two values 40 or 60. It is then convenient to introduce a change in the interpretation of θ. Now when $\theta = 1$, it will mean that the common pdf of X_1, X_2, X_3, Y is $0.02e^{-0.02x}$, and when $\theta = 2$, it will mean that the common pdf is $0.04e^{-0.04x}$. In this case the convex set C is given by Fig. 5.6. The computations giving the boundary of C in this case are tedious, and will be discussed in Sec. 5.12. Each point on the boundary of C corresponds to a decision rule which does not use randomization. There is only one point corresponding to an admissible decision rule.

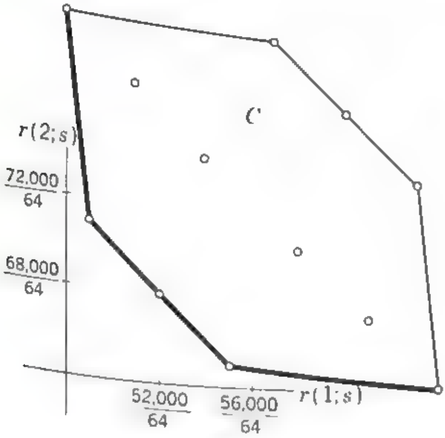


Fig. 5.5

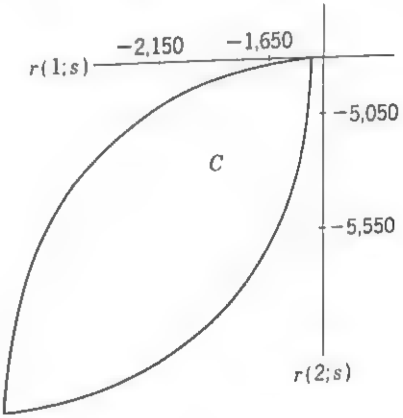


Fig. 5.6

5.8. Bayes Decision Rules. Throughout this section, we continue our assumption that there are h different possible joint distributions. If $b(1), b(2), \ldots, b(h)$ is a set of nonnegative numbers with $b(1) + b(2) + \cdots + b(h) = 1$, then a decision rule s is called a "Bayes decision rule relative to $b(1), b(2), \ldots, b(h)$" if

$$\sum_{\theta=1}^{h} b(\theta)r(\theta;s) \leq \sum_{\theta=1}^{h} b(\theta)r(\theta;t)$$

for each and every decision rule t.

If a decision rule s is called simply a "Bayes decision rule," it means that there is some (unspecified) set of nonnegative numbers $b(1), b(2), \ldots, b(h)$, with $b(1) + b(2) + \cdots + b(h) = 1$, such that s is a Bayes decision rule relative to $b(1), b(2), \ldots, b(h)$.

The theorem of Sec. 5.7 tells us that any admissible decision rule must be a Bayes decision rule. This means that in our search for admissible decision rules, we can limit attention to the Bayes decision rules.

In some statistical problems, there are Bayes decision rules which are inadmissible, which does not contradict the fact that any admissible decision rule is a Bayes decision rule. For example, in the first illustrative example of Sec. 5.7, the convex set C is a horizontal line segment, and it is easily seen that *every* decision rule is a Bayes decision rule relative to 0,1. However, there is only one admissible decision rule. Therefore it may be wondered why we bother to pay attention to Bayes decision rules, which may include among them some inadmissible decision rules, when what we would really like to find are admissible decision rules. The answer is that it is so simple to find the Bayes decision rules that it is a useful first step in our search for the admissible decision rules.

If s is a Bayes decision rule relative to $b(1), \ldots, b(h)$ and $b(\theta) > 0$ for $\theta = 1, \ldots, h$, then s must be admissible. This is proved as follows. Suppose s were not admissible. Then there would be a decision rule t with $r(\theta;t) \leqslant r(\theta;s)$ for $\theta = 1, \ldots, h$, with $r(\theta;t) < r(\theta;s)$ for at least one value of θ, say, for $\theta = j$. But then

$$\sum_{\theta=1}^{h} b(\theta)r(\theta;s) - \sum_{\theta=1}^{h} b(\theta)r(\theta;t) = \sum_{\theta=1}^{h} b(\theta)[r(\theta;s) - r(\theta;t)] \geqslant b(j)[r(j;s) - r(j;t)] > 0$$

and therefore

$$\sum_{\theta=1}^{h} b(\theta)r(\theta;s) > \sum_{\theta=1}^{h} b(\theta)r(\theta;t),$$

which contradicts the fact that s is a Bayes decision rule relative to $b(1), \ldots, b(h)$. This contradiction proves that s is admissible.

5.9. The Construction of Bayes Decision Rules When There Is a Finite Number of Distributions and a Finite Number of Possible Decisions. First we consider the case where each possible distribution allows us to list the possible values of the chance variables. Then, for the decision rule s with decision probabilities $s(D;x)$, we have from Sec. 5.6 that

$$r(\theta;s) = \sum_{x}\sum_{y}\sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)s(D;x)$$

Suppose we are given a set of h nonnegative numbers $b(1), \ldots, b(h)$, with $b(1) + \cdots + b(h) = 1$, and we want to find a decision rule s that is a Bayes decision rule relative to $b(1), \ldots, b(h)$. Then s must be chosen to make $\sum_{\theta=1}^{h} b(\theta)\, r(\theta;s)$ as small as possible. But

$$\sum_{\theta=1}^{h} b(\theta)r(\theta;s) = \sum_{x}\sum_{D=1}^{L} s(D;x)\left[\sum_{\theta=1}^{h}\sum_{y} b(\theta)W(y;d;x)f(x,y;\theta)\right]$$

Let us denote the expression $\sum_{\theta=1}^{h}\sum_{y} b(\theta)\; W(y;D;x)\, f(x,y;\theta)$ by $K(D;x)$. Then s must be chosen to make $\sum_{x}\sum_{D=1}^{L} s(D;x)\, K(D;x)$ as small as possible. This will be done if for each x, the quantities $s(1;x), \ldots, s(L;x)$ are set so as to minimize $\sum_{D=1}^{L} s(D;x)\, K(D;x)$. Since $s(1;x), \ldots, s(L;x)$ are nonnegative and add to unity, it is clear that the minimum will be achieved if and only if $s(D;x)$ is set equal to zero for every D for which $K(D;x)$ is greater than the smallest of the quantities $K(1;x), \ldots, K(L;x)$. In other words, for a given x we should never choose a D that did not minimize $K(D;x)$.

From our discussion, we see that different decision rules may be Bayes decision rules relative to the same $b(1), \ldots, b(h)$. This happens when for some x, $K(D;x)$ is minimized for more than one value of D. If i and j are two different integers with $K(i;x) = K(j;x) = \min [K(1;x), \ldots, K(L;x)]$, our construction above says nothing about the relative sizes of $s(i;x)$ and $s(j;x)$, though of course

$$s(D;x) \geq 0 \qquad \text{and} \qquad \sum_{D=1}^{L} s(D;x) = 1$$

are conditions that must always be satisfied.

As an illustration, we take the second example of Sec. 5.7, in which there are two possible distributions, given by proportions of defectives of $\frac{1}{4}$, $\frac{3}{4}$, respectively. We shall find a Bayes decision rule relative to $\frac{1}{2}$, $\frac{1}{2}$. To do this, we compute $K(1;x)$ and $K(2;x)$ for all four possible values of x. The computations are as follows:

$$\begin{aligned}
K(1;0,0) &= \tfrac{1}{2}[(500)(\tfrac{3}{4})^3 + (1{,}500)(\tfrac{1}{4})(\tfrac{3}{4})^2] \\
&\quad + \tfrac{1}{2}[(500)(\tfrac{1}{4})^3 + (1{,}500)(\tfrac{3}{4})(\tfrac{1}{4})^2] = 32{,}000/128 \\
K(2;0,0) &= \tfrac{1}{2}[(1{,}000)(\tfrac{3}{4})^3 + (1{,}000)(\tfrac{1}{4})(\tfrac{3}{4})^2] \\
&\quad + \tfrac{1}{2}[(1{,}000)(\tfrac{1}{4})^3 + (1{,}000)(\tfrac{3}{4})(\tfrac{1}{4})^2] = 40{,}000/128 \\
K(1;0,1) &= K(1;1,0) = \tfrac{1}{2}[(500)(\tfrac{1}{4})(\tfrac{3}{4})^2 + (1{,}500)(\tfrac{1}{4})^2(\tfrac{3}{4})] \\
&\quad + \tfrac{1}{2}[(500)(\tfrac{3}{4})(\tfrac{1}{4})^2 + (1{,}500)(\tfrac{3}{4})^2(\tfrac{1}{4})] = 24{,}000/128 \\
K(2;0,1) &= K(2;1,0) = \tfrac{1}{2}[(1{,}000)(\tfrac{1}{4})(\tfrac{3}{4})^2 + (1{,}000)(\tfrac{1}{4})^2(\tfrac{3}{4})] \\
&\quad + \tfrac{1}{2}[(1{,}000)(\tfrac{3}{4})(\tfrac{1}{4})^2 + (1{,}000)(\tfrac{3}{4})^2(\tfrac{1}{4})] = 24{,}000/128 \\
K(1;1,1) &= \tfrac{1}{2}[(500)(\tfrac{1}{4})^2(\tfrac{3}{4}) + (1{,}500)(\tfrac{1}{4})^3] \\
&\quad + \tfrac{1}{2}[(500)(\tfrac{3}{4})^2(\tfrac{1}{4}) + (1{,}500)(\tfrac{3}{4})^3] = 48{,}000/128 \\
K(2;1,1) &= \tfrac{1}{2}[(1{,}000)(\tfrac{1}{4})^2(\tfrac{3}{4}) + (1{,}000)(\tfrac{1}{4})^3] \\
&\quad + \tfrac{1}{2}[(1{,}000)(\tfrac{3}{4})^2(\tfrac{1}{4}) + (1{,}000)(\tfrac{3}{4})^3] = 40{,}000/128
\end{aligned}$$

Since $K(1;0,0) < K(2;0,0)$, a Bayes decision rule relative to $\frac{1}{2}$, $\frac{1}{2}$ surely chooses decision 1 when $X_1 = X_2 = 0$. Similarly, a Bayes decision rule relative to $\frac{1}{2}$, $\frac{1}{2}$ surely chooses decision 2 when $X_1 = X_2 = 1$. Since $K(1;1,0) = K(2;1,0)$ and $K(1;0,1) = K(2;0,1)$, a Bayes decision rule s relative to $\frac{1}{2}$, $\frac{1}{2}$ can assign any value between 0 and 1 to $s(1;1,0)$ and $s(1;0,1)$. In summary, a decision rule s is a Bayes decision rule relative to $\frac{1}{2}$, $\frac{1}{2}$ if and only if $s(1;0,0) = 1$ and $s(1;1,1) = 0$. For such a decision rule s,

$$\begin{aligned}
r(1;s) &= 55{,}000/64 - (3{,}000/64)[s(1;0,1) + s(1;1,0)] \\
r(2;s) &= 65{,}000/64 + (3{,}000/64)[s(1;0,1) + s(1;1,0)]
\end{aligned}$$

As $s(1;0,1)$ and $s(1;1,0)$ vary between 0 and 1, the points $[r(1;s), r(2;s)]$ give the line segment joining (49,000/64, 71,000/64) and (55,000/64, 65,000/64). (See Fig. 5.5 in Sec. 5.7.)

As another illustrative example, we find a Bayes decision rule relative to $\frac{2}{5}$, $\frac{3}{5}$ for the preceding problem. We compute $K(1;x)$ and $K(2;x)$ as follows:

$$K(1;0,0) = \tfrac{2}{5}[(500)(\tfrac{3}{4})^3 + (1{,}500)(\tfrac{1}{4})(\tfrac{3}{4})^2] + \tfrac{3}{5}[(500)(\tfrac{1}{4})^3 + (1{,}500)(\tfrac{3}{4})(\tfrac{1}{4})^2] = 13{,}800/64$$

$$K(2;0,0) = \tfrac{2}{5}[(1{,}000)(\tfrac{3}{4})^3 + (1{,}000)(\tfrac{1}{4})(\tfrac{3}{4})^2] + \tfrac{3}{5}[(1{,}000)(\tfrac{1}{4})^3 + (1{,}000)(\tfrac{3}{4})(\tfrac{1}{4})^2] = 16{,}800/64$$

$$K(1;0,1) = K(1;1,0) = 12{,}600/64$$

$$K(2;0,1) = K(2;1,0) = 12{,}000/64$$

$$K(1;1,1) = 28{,}200/64$$

$$K(2;1,1) = 23{,}200/64$$

There is only one Bayes decision rule relative to $\frac{2}{5}$, $\frac{3}{5}$: it is the decision rule s with $s(1;0,0) = 1$, $s(1;0,1) = s(1;1,0) = s(1;1,1) = 0$. For this decision rule s, $r(1;s) = 55{,}000/64$, $r(2;s) = 65{,}000/64$.

Next we consider the case where each possible distribution has a joint pdf for $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. $f(x,y;\theta)$ denotes the joint pdf for the θth distribution in our list. Using the formula given in Sec. 5.6,

$$r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{D=1}^{L} W(y;D;x) f(x,y;\theta) s(D;x)\, dx_1 \cdots dx_m\, dy_1 \cdots dy_n$$

and

$$\sum_{\theta=1}^{h} b(\theta) r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{D=1}^{L} s(D;x)$$

$$\times \left[\sum_{\theta=1}^{h} b(\theta) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} W(y;D;x) f(x,y;\theta)\, dy_1 \cdots dy_n\right] dx_1 \cdots dx_m$$

Denoting

$$\sum_{\theta=1}^{h} b(\theta) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} W(y;D;x) f(x,y;\theta)\, dy_1 \cdots dy_n$$

by $K(D;x)$, we have

$$\sum_{\theta=1}^{h} b(\theta) r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{D=1}^{L} s(D;x) K(D;x)\, dx_1 \cdots dx_m$$

In order for s to be a Bayes decision rule relative to $b(1), b(2), \ldots, b(h)$, we must choose the values $s(D;x)$ to minimize $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{D=1}^{L} s(D;x) \times K(D;x)\, dx_1 \cdots dx_m$. This will be done if for each x, the quantities $s(1;x), \ldots, s(L;x)$ are set so as to minimize $\sum_{D=1}^{L} s(D;x)\, K(D;x)$. By the same reasoning already used, the minimum will be achieved if $s(D;x)$ is set equal to zero for every D for which $K(D;x)$ is greater than the smallest of the quantities $K(1;x), \ldots, K(L;x)$.

As a numerical example, we simplify Example 2 of Sec. 5.6 by assuming that D can take only the values 20, 40, and 60 and no others and that the value of the unknown parameter is known to be one of the values 0.01, 0.03, 0.05. To bring the notation into line with the notation we have been using, where D ranged from 1 to L and θ from 1 to h, we should really relabel our decisions and distributions. However, we shall not bother to do this, and no confusion will result. Suppose we want to find a Bayes decision rule s relative to $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$; that is, we want to find a decision rule s that will minimize $(\frac{1}{4})r(0.01;s) + (\frac{1}{2})r(0.03;s) + (\frac{1}{4})r(0.05;s)$. We find

$$K(D;x) = [(\tfrac{1}{4})(0.01)^3 e^{-0.01(x_1+x_2+x_3+D)} + (\tfrac{1}{2})(0.03)^3 e^{-0.03(x_1+x_2+x_3+D)}$$
$$+ (\tfrac{1}{4})(0.05)^3 e^{-0.05(x_1+x_2+x_3+D)}](-10{,}000 - 100D)$$

Clearly, $K(D;x)$ depends on x_1, x_2, x_3 only through the sum $x_1 + x_2 + x_3$. Given any specific value for $x_1 + x_2 + x_3$, we can compute $K(20;x)$, $K(40;x)$, $K(60;x)$ and take the appropriate action. Thus, if $x_1 + x_2 + x_3$ equals 80, $K(20;x) = -0.0117$, $K(40;x) = -0.0073$, $K(60;x) = -0.00469$, and s must have $s(20;x) = 1$, so that if $X_1 + X_2 + X_3 = 80$, the decision $D = 20$ is certainly chosen.

5.10. The Construction of Bayes Decision Rules When There Is an Infinite Number of Possible Distributions and a Finite Number of Possible Decisions. Our definition of admissible decision rule covers the case of an infinite number of possible distributions, but our definition of Bayes decision rule covers only the case of a finite number of possible distributions. Our first task is to extend the definition of Bayes decision rule to the case of an infinite number of possible distributions.

In the case of a finite number of possible distributions, a Bayes decision rule s relative to $b(1), \ldots, b(h)$ is a decision rule that makes $\sum_{\theta=1}^{h} b(\theta) r(\theta;s)$ as small as possible. We note that if θ were a chance variable with the distribution

Possible values	1	2	$\cdots$	h
Probabilities	$b(1)$	$b(2)$	$\cdots$	$b(h)$

then $\sum_{\theta=1}^{h} b(\theta)r(\theta;s)$ would be $E\{r(\theta;s)\}$ for each decision rule s. Of course, we do *not* consider θ to be a chance variable, but the interpretation of $\sum_{\theta=1}^{h} b(\theta)r(\theta;s)$ as the expected value of a function of a pretended chance variable θ allows us to extend the definition of a Bayes decision rule to the case of an infinite number of possible distributions.

Suppose $B(\theta)$ is a given cdf for a chance variable θ, which assigns all the probability to the possible values of θ in our decision problem. For any decision rule s, we denote by $R(s;B(\theta))$ the expected value $r(\theta;s)$ *would* have *if* θ were a chance variable with cdf $B(\theta)$. A decision rule s is called a "Bayes decision rule relative to $B(\theta)$" if $R(s;B(\theta)) \leq R(t;B(\theta))$ for each and every decision rule t. This definition covers the case of a finite number of possible distributions, as well as the case of an infinite number of possible distributions, and coincides with the definition previously given for the case of a finite number of possible distributions. A cdf used for the purpose of defining a Bayes decision rule, as $B(\theta)$ was used, is called an "a priori distribution." Again we emphasize the fact that θ is not a chance variable, but an unknown constant. The introduction of the cdf $B(\theta)$ is just a technical device to enable us to extend the definition of a Bayes decision rule to the case of an infinite number of possible distributions.

If $B(\theta)$ can be differentiated to give a pdf $b(\theta)$, then $R(s;B(\theta)) = \int r(\theta;s)b(\theta)\,d\theta$. Let us assume that we are dealing with a problem where $r(\theta;s)$ is a continuous function of θ, for each decision rule s. Then if the pdf $b(\theta)$ is positive for all possible values of θ, and if s is a Bayes decision rule relative to $B(\theta)$, s must be admissible. To prove this, suppose s were not admissible. Then there would be a decision rule t, with $r(\theta;t) \leq r(\theta;s)$ for all θ, with strict inequality for at least one value of θ, say, for $\bar{\theta}$. Thus $r(\bar{\theta};t) < r(\bar{\theta};s)$. But $r(\theta;s)$ and $r(\theta;t)$ are continuous functions of θ, and therefore there must be values A, B, with $A < B$ and $\bar{\theta}$ contained between A and B, such that $r(\theta;t) < r(\theta;s)$ for *all* θ in the interval from A to B. Since

$$R(t;B(\theta)) - R(s;B(\theta)) \leq \int_A^B [r(\theta;t) - r(\theta;s)]b(\theta)\,d\theta < 0$$

we get $R(t;B(\theta)) < R(s;B(\theta))$, which is a contradiction, since s is a Bayes decision rule relative to $B(\theta)$. This contradiction proves that s is admissible. [The nonexistence of $b(\theta)$ at a finite number of points does not affect the argument.]

The preceding paragraph shows that in the case of an infinite number of possible distributions, the class of all Bayes decision rules contains many admissible decision rules. But are *all* the admissible decision rules contained among the Bayes decision rules, as in the case of a finite number of possible distributions? The answer is, in general, in the negative: some admissible decision rules are not Bayes decision rules. To examine this situation further, suppose $s_1, s_2, \ldots$ is an infinite sequence of Bayes decision rules. A decision rule s is called the "limit of the sequence $s_1, s_2, \ldots$" if

$$r(\theta;s) = \lim_{j\to\infty} r(\theta;s_j)$$

for all possible values of θ in the problem. Then we have the following theorem for the case of an infinite number of possible distributions: *Each admissible decision rule is either a Bayes decision rule, or else is a limit of a sequence of Bayes decision rules.* We shall not prove this theorem.

For a given a priori distribution $B(\theta)$, the construction of a Bayes decision rule relative to $B(\theta)$ is a straightforward matter. For example, suppose $B(\theta)$ has a pdf $b(\theta)$, and for each θ, the joint distribution of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ allows the possible values to be listed. Then

$$R(s;B(\theta)) = \int\left[\sum_x \sum_y \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)s(D;x)\right]b(\theta)\,d\theta$$

$$= \sum_x \sum_{D=1}^{L} s(D;x)K(D;x)$$

where
$$K(D;x) = \int \sum_y W(y;D;x)f(x,y;\theta)b(\theta)\,d\theta$$

Then s is a Bayes decision rule relative to $B(\theta)$ if for each x, $s(D;x)$ is set equal to zero for every D for which $K(D;x)$ is greater than the smallest of the quantities $K(1;x), \ldots, K(L;x)$. If for each θ, the joint distribution of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ has a pdf $f(x,y;\theta)$, then we define $K(D;x)$ as $\int [\int \cdots \int W(y;D;x)f(x,y;\theta)\, dy_1 \cdots dy_n]b(\theta)\,d\theta$ and use it the same way as above.

As a numerical example, we take Example 1 of Sec. 5.6 and find a Bayes decision rule relative to the cdf $B(\theta)$ defined as follows:

$$B(\theta) = 0 \qquad \text{for } \theta \leqslant 0$$

$$B(\theta) = \theta \qquad \text{for } 0 \leqslant \theta \leqslant 1$$

$$B(\theta) = 1 \qquad \text{for } \theta > 1$$

The pdf $b(\theta)$ corresponding to $B(\theta)$ is

$$b(\theta) = 1 \qquad \text{for } 0 < \theta < 1$$

$$b(\theta) = 0 \qquad \text{for } \theta < 0 \text{ or } \theta > 1$$

Computing $K(D;x)$, we get

$$K(1;0,0) = \int_0^1 [500(1-\theta)^3 + 1{,}500\theta(1-\theta)^2]\,d\theta = 250$$

$$K(2;0,0) = \int_0^1 [1{,}000(1-\theta)^3 + 1{,}000\theta(1-\theta)^2]\,d\theta = 1{,}000/3$$

$$K(1;0,1) = K(1;1,0) = \int_0^1 [500\theta(1-\theta)^2 + 1{,}500\theta^2(1-\theta)]\,d\theta = 2{,}000/12$$

$$K(2;0,1) = K(2;1,0) = \int_0^1 [1{,}000\theta(1-\theta)^2 + 1{,}000\theta^2(1-\theta)]\,d\theta = 2{,}000/12$$

$$K(1;1,1) = \int_0^1 [500\theta^2(1-\theta) + 1{,}500\theta^3]\,d\theta = 5{,}000/12$$

$$K(2;1,1) = \int_0^1 [1{,}000\theta^2(1-\theta) + 1{,}000\theta^3]\,d\theta = 1{,}000/3$$

Thus any decision rule s with $s(1;0,0) = 1$ and $s(1;1,1) = 0$ is a Bayes decision rule relative to the given $B(\theta)$. Since it is easily verified that $r(\theta;s)$ in this problem is a continuous function of θ for each given s, and since our $b(\theta)$ is positive for all possible θ (with the inconsequential exception of the points $\theta = 0$, $\theta = 1$), any decision rule s which is Bayes relative to the given $B(\theta)$ is admissible. In particular, the decision rule s_1 of Example 1 of Sec. 5.6 is admissible.

In the discussion above, we have been implicitly assuming that the different possible joint distributions are given by the variation of a single parameter θ. However, in many important problems, the possible joint distributions are given by the variation of more than one parameter. For example, the possible distributions may be all possible normal distributions, and two separate parameters are necessary to specify a normal distribution. In such a case, we use as an a priori distribution a joint cdf $B(\theta_1,\theta_2)$ for the parameters θ_1, θ_2. All the definitions and computations remain essentially the same.

5.11. The Construction of Bayes Decision Rules When There Is an Infinite Number of Possible Decisions. When there is an infinite number of possible decisions, we compute $K(D;x)$ exactly as we did above. Then for each x, we set $s(D;x) = 0$ unless $K(D;x) = \min_{\Delta} K(\Delta;x)$.

As a numerical example, we take Example 2 of Sec. 5.6, and we find a Bayes decision rule s relative to the a priori distribution $B(\theta) = 1 - e^{-\theta}$ for $\theta \geq 0$. This a priori cdf has pdf $b(\theta) = e^{-\theta}$ for $\theta > 0$. Then we have

$$\begin{aligned} K(D;x) &= \int_0^\infty \left\{ \int_0^D 0\,dy + \int_D^\infty (10{,}000 + 100D)\theta^4 \right. \\ &\qquad \left. \times \exp\,[-\theta(x_1 + x_2 + x_3 + y)]\,dy \right\} e^{-\theta}\,d\theta \\ &= -6(10{,}000 + 100D)(x_1 + x_2 + x_3 + D + 1)^{-4} \end{aligned}$$

A simple calculation shows that $K(D;x)$ is minimized when $D = (1/3)(x_1 + x_2 + x_3) - 133$. But of course D can never be negative, and this must be taken into account. Then we get as a Bayes decision rule relative to the given $B(\theta)$: Choose D equal to $(1/3)(X_1 + X_2 + X_3) - 133$ if this quantity is positive; otherwise choose $D = 0$. This decision rule is admissible, since $b(\theta) > 0$ for all $\theta > 0$.

5.12. The Geometry of Bayes Decision Rules. In Sec. 5.7 we stated that an easy way to sketch the convex set C discussed there was to find the points corresponding to decision rules not using randomization, and then draw line segments between pairs of these points to get the boundary of C. We shall prove this now. Although the proof is carried out for the case of two-dimensional sets C, it is true in general that the boundary of C is determined by the points corresponding to decision rules not using randomization.

For given values A, B, not both equal to zero, let $g_1(A,B)$ denote the smallest value of g such that the line $Ar(1;s) + Br(2;s) = g$ contains points of the convex set C, and let $g_2(A,B)$ denote the largest value of g such that the line $Ar(1;s) + Br(2;s) = g$ contains points of the convex set C. For example, in Example 2 of Sec. 5.7, $g_1(1,1) = 120{,}000/64$, $g_2(1,1) = 136{,}000/64$, $g_1(-1,1) = 0$, $g_2(-1,1) = 32{,}000/64$. All points of C which are on the lines $Ar(1;s) + Br(2;s) = g_1(A,B)$, $Ar(1;s) + Br(2;s) = g_2(A,B)$ are on the boundary of C. If A, B are both nonnegative, then the points of C on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$ represent Bayes decision rules relative to $A/(A + B)$, $B/(A + B)$.

To find the points of C on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$, we compute

$$K(D;x) = A\sum_{y} W(y;D;x)f(x,y;1) + B\sum_{y} W(y;D;x)f(x,y;2)$$

and set $s(D;x) = 0$ unless

$$K(D;x) = \min_{\Delta} K(\Delta;x)$$

The resulting decision rule s corresponds to a point on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$. We note that there is always such an s which does not use randomization.

To find the points of C on the line $Ar(1;s) + Br(2;s) = g_2(A,B)$, we proceed as in the preceding paragraph, except that we set $s(D;x) = 0$ unless

$$K(D;x) = \max_{\Delta} K(\Delta;x)$$

The resulting decision rule corresponds to a point on the line $Ar(1;s) + Br(2;s) = g_2(A,B)$. We note that there is always such a decision rule which does not use randomization.

Now we can show that each point on the boundary of C either represents a decision rule not using randomization, or else lies on a line segment joining two points of C representing decision rules not using randomization. Suppose this were not so. Denote by C' the convex set we get by plotting all the points corresponding to decision rules not using randomization, and then drawing line segments between pairs of these points in such a way that each of the points is either on one of the line segments or else contained inside the boundary formed by the line segments, and the line segments enclose a convex set. Then we have Fig. 5.7, where p is a point on the boundary of C. But then there is an A, B such that the point p is on one of the lines $Ar(1;s) + Br(2;s) = g_1(A,B)$, $Ar(1;s) + Br(2;s) = g_2(A,B)$, while no point of C' is on the line. However, this is a contradiction, since the line must contain a point corresponding to a decision rule not using randomization; that is, it must contain a point of C'. This proves that each point on the boundary of C either represents a decision rule not using randomization, or else lies on a line segment joining two points of C representing decision rules not using randomization.

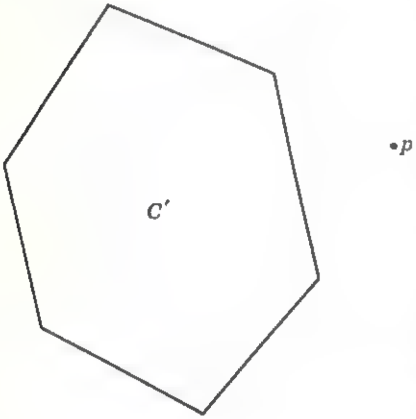


Fig. 5.7

In general, it is true that the boundary of C (and therefore C itself) is completely determined by the decision rules not using randomization.

For the third example of Sec. 5.7, we sketched the convex set C without showing the computations. Now we are in a position to go through the computations. We fix A, B and find the decision rules which lie on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$. We have

$$K(40;x) = -A\int_{40}^{\infty}(14{,}000)(0.02)^4 \exp\left[-0.02(x_1 + x_2 + x_3 + y)\right] dy$$

$$- B\int_{40}^{\infty}(14{,}000)(0.04)^4 \exp\left[-0.04(x_1 + x_2 + x_3 + y)\right] dy$$

$$K(60;x) = -A\int_{60}^{\infty}(16{,}000)(0.02)^4 \exp\left[-0.02(x_1 + x_2 + x_3 + y)\right] dy$$

$$- B\int_{60}^{\infty}(16{,}000)(0.04)^4 \exp\left[-0.04(x_1 + x_2 + x_3 + y)\right] dy$$

After the integrations are performed, it turns out that if A and B are both positive, then $K(40;x) < K(60;x)$ for all possible values of x_1, x_2, x_3, so that if A, B are positive, the only decision rule on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$ is the decision rule for which $s(40;x) = 1$ for all x. For this decision rule, $r(1;s)$ is approximately $-6{,}290$, $r(2;s)$ is approximately $-2{,}830$.

Denote

$$50 \log \frac{-B}{A} + 50 \log \left[\frac{8e^{-0.8} - (64/7)e^{-1.6}}{1 - (8/7)e^{-0.4}} \right]$$

by w. If A is positive and B is negative, $K(40;x) < K(60;x)$ if and only if $x_1 + x_2 + x_3 > w$. Thus if A is positive and B is negative, the only decision rule on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$ is the decision rule for which $s(40;x) = 1$ if $x_1 + x_2 + x_3 > w$, and $s(40;x) = 0$ if $x_1 + x_2 + x_3 < w$. For this decision rule,

$$r(1;s) = -4{,}810 - 1{,}480e^{-0.02w}[1 + 0.02w + (\tfrac{1}{2})(0.02w)^2]$$
$$r(2;s) = -1{,}450 - 1{,}380e^{-0.04w}[1 + 0.04w + (\tfrac{1}{2})(0.04w)^2]$$

If A is negative and B is positive, $K(40;x) < K(60;x)$ if and only if $x_1 + x_2 + x_3 < w$. In this case, the only decision rule on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$ is the decision rule for which $s(40;x) = 1$ if $x_1 + x_2 + x_3 < w$, and $s(40;x) = 0$ if $x_1 + x_2 + x_3 > w$. For this decision rule, $r(1;s) = -6{,}290 + 1{,}480e^{-0.02w}[1 + 0.02w + (\tfrac{1}{2})(0.02w)^2]$, $r(2;s) = -2{,}830 + 1{,}380e^{-0.04w}[1 + 0.04w + (\tfrac{1}{2})(0.04w)^2]$. If A and B are both negative, there is only one decision rule on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$, and it has $r(1;s) = -4{,}810$, $r(2;s) = -1{,}450$. Thus we have found all the boundary points of C on the line $Ar(1;s) + Br(2;s) = g_1(A,B)$. The boundary points of C on the line $Ar(1;s) + Br(2;s) = g_2(A,B)$ are already included in the ones we have found, because $g_1(-A,-B) = -g_2(A,B)$.

Note that, in the preceding example, every point on the boundary of C corresponded to a decision rule not using randomization. This illustrates the following theorem: *In any decision problem where there is a finite number of possible joint distributions and each of the distributions has a pdf, then corresponding to any point on the boundary of the convex set C there is a decision rule not using randomization.*

We shall not prove the theorem just stated, but we point out an important consequence. If s is any admissible decision rule, s corresponds to a point Q on the boundary of C. But by the theorem just stated, there is a decision rule t not using randomization which corresponds to the point Q. This means that $r(\theta;s) = r(\theta;t)$ for all θ, which in turn means that s and t are equivalent from our point of view. Thus we need never use a decision rule which uses randomization, since there is always a decision rule not using randomization which is as good. (We emphasize that this last statement applies to a problem with a finite number of possible distributions, each having a pdf.)

5.13. Sufficiency. Suppose that we have a decision problem in which there are u functions of $x_1, \ldots, x_m$, denoted by $z_1, \ldots, z_u$, with the following two properties:

1. There are functions $A(x_1, \ldots, x_m)$, $g(z_1, \ldots, z_u, y_1, \ldots, y_n; \theta)$, never negative, such that for all θ, $f(x_1, \ldots, x_m, y_1, \ldots, y_n; \theta) = A(x_1, \ldots, x_m)g(z_1, \ldots, z_u, y_1, \ldots, y_n; \theta)$, identically in the x's and y's.

2. There is a function $\overline{W}(y_1, \ldots, y_n; D; z_1, \ldots, z_u)$ such that $W(y_1, \ldots, y_n; D; x_1, \ldots, x_m) = \overline{W}(y_1, \ldots, y_n; D; z_1, \ldots, z_u)$, identically in the arguments. Then we shall show that any admissible decision rule can be based on a knowledge of the values of $z_1, \ldots, z_u$ alone, with no necessity for knowing the individual values of $x_1, \ldots, x_m$.

Before showing this, we give two examples. First we note that property 2 is automatically satisfied in any problem in which the loss function depends only on $y_1, \ldots, y_n$, D, and not on $x_1, \ldots, x_m$. Turning to Example 1 of Sec. 5.6, we note that $f(x_1,x_2,y;\theta)$ can be written as $\theta^{x_1+x_2+y}(1-\theta)^{3-(x_1+x_2+y)}$. Setting $z = x_1 + x_2$, we have $f(x_1,x_2,y;\theta) = \theta^{z+y}(1-\theta)^{3-(z+y)}$. Thus property 1 is satisfied with $u = 1$, $z_1 = x_1 + x_2$, $A(x_1,x_2) = 1$. Property 2 is automatically satisfied, since the loss function W does not depend on x_1, x_2.

As another example, we turn to Example 2 of Sec. 5.6. There $f(x_1,x_2,x_3,y;\theta) = \theta^4 \exp[-\theta(x_1 + x_2 + x_3 + y)]$ if x_1, x_2, x_3, y are all positive. Setting $z = x_1 + x_2 + x_3$, we see that property 1 is satisfied with $u = 1$, $z_1 = x_1 + x_2 + x_3$, $A(x_1,x_2,x_3) = 1$. Property 2 is also satisfied, since the loss is a function only of y and D.

Now we turn to the proof that any admissible decision rule can be based purely on a knowledge of $z_1, \ldots, z_u$. Suppose we want to construct a Bayes decision rule relative to $B(\theta)$, where $B(\theta)$ has pdf $b(\theta)$. Then $K(D;x) = \int b(\theta)[\int \cdots \int W(y;D;x)f(x,y;\theta)\,dy_1 \cdots dy_n]\,d\theta = A(x_1, \ldots, x_m) \int b(\theta)\,[\int \cdots \int \overline{W}(y_1, \ldots, y_n;\ D;\ z_1, \ldots, z_u)g(z_1, \ldots, z_u, y_1, \ldots, y_n;\ \theta)\ dy_1 \cdots dy_n]\,d\theta$. We denote $\int b(\theta)[\int \cdots \int \overline{W}(y_1, \ldots, y_n;\ D;\ z_1, \ldots, z_u)\ g(z_1, \ldots, z_u, y_1, \ldots y_n;\ \theta)\,dy_1 \cdots dy_n]\,d\theta$ by $\overline{K}(D;z)$. Thus $K(D;x) = A(x_1, \ldots, x_m)\overline{K}(D;z)$, and this means that

$$K(D;x) = \min_{\Delta} K(\Delta;x)$$

if and only if

$$\overline{K}(D;z) = \min_{\Delta} \overline{K}(\Delta;z)$$

This means that we can construct any decision rule not using randomization just as well from a knowledge of the values of $z_1, \ldots, z_u$ as from a knowledge of the values of $x_1, \ldots, x_m$, since the D that minimizes

$K(D;x)$ also minimizes $\bar{K}(D;z)$. But we have seen in the preceding section that the decision rules not using randomization determine the whole convex set C that is available to us. This implies that a person who knows only the values $z_1, \ldots, z_u$ can construct a decision rule identical (as far as expected losses are concerned) with any decision rule that can be constructed by a person who knows all the values $x_1, \ldots, x_m$. For this reason, it is said that "$z_1, \ldots, z_u$ are sufficient for the decision problem."

If u is much smaller than m, as is often the case, it is convenient to restate the whole problem in terms of $z_1, \ldots, z_u$. This is done as follows. Let $Z_1, \ldots, Z_u$ be the same functions of $X_1, \ldots, X_m$ as $z_1, \ldots, z_u$ are of $x_1, \ldots, x_m$. Then the original decision problem can be restated as a decision problem in which we observe $Z_1, \ldots, Z_u$ and then make a decision. Thus, in our first example above, $Z = X_1 + X_2$ has a binomial distribution with parameters 2, θ, so the possible joint distributions of Z, Y are given by

		Possible values of Z		
		0	1	2
Possible values of Y	0	$(1-\theta)^3$	$2\theta(1-\theta)^2$	$\theta^2(1-\theta)$
	1	$\theta(1-\theta)^2$	$2\theta^2(1-\theta)$	θ^3

as θ varies from 0 to 1. The loss function $W(y;D)$ remains the same.

In our second example above, $Z = X_1 + X_2 + X_3$, and by a calculation similar to that used in the last example of Sec. 3.11, we find that the possible joint pdf's of Z, Y are $(\frac{1}{2})\theta^4 z^2 e^{-\theta(z+y)}$ for y, z both positive, θ varying between 0 and ∞. The loss function $W(y;D)$ remains the same.

As a third example, suppose a company has to decide how much of a perishable commodity should be stocked to meet the demand of the coming sales period. Since this demand is the sum of many independent individual demands, it is assumed that the demand has a normal distribution (Sec. 4.7) with unknown mean θ_1 and unknown standard deviation θ_2. The net profit on each unit of the commodity sold is p_1 dollars; the net loss on each unit of the commodity unsold at the end of the period is p_2 dollars. Before deciding, the company will observe the demands for the commodity in m regions similar to the region it serves, the demands in these m regions being assumed independent, each with the same distribution as the demand for the coming sales period. Assuming that the cost of observing the m regions is negligible, the decision problem has the following structure. D is the amount of the commodity that will be stocked and can be any positive number. Y is the demand that will be observed in the coming period. $X_1, \ldots, X_m$ are the demands that will be observed in the m regions. The loss depends only on Y and D, and is given as follows:

$$W(Y;D) = -p_1 D \qquad \text{if } Y \geqslant D$$
$$W(Y;D) = -p_1 Y + p_2(D - Y) \qquad \text{if } Y < D$$

The possible joint pdf's are given by

$$(\theta_2\sqrt{2\pi})^{-m-1} \exp\left[-\frac{1}{2\theta_2^{\,2}}\sum_{i=1}^{m}(x_i - \theta_1)^2 - \frac{1}{2\theta_2^{\,2}}(y - \theta_1)^2\right]$$

as θ_1, θ_2 vary, with θ_2 always positive. Define z_1 as $(1/m)(x_1 + \cdots + x_m)$ and z_2 as $(1/m)\sum_{i=1}^{m}(x_i - z_1)^2$. We have

$$\begin{aligned}\sum_{i=1}^{m}(x_i - \theta_1)^2 &= \sum_{i=1}^{m}[(x_i - z_1) + (z_1 - \theta_1)]^2 \\ &= \sum_{i=1}^{m}(x_i - z_1)^2 + m(z_1 - \theta_1)^2 + 2\sum_{i=1}^{m}(x_i - z_1)(z_1 - \theta_1)\end{aligned}$$

But

$$\sum_{i=1}^{m}(x_i - z_1)(z_1 - \theta_1) = (z_1 - \theta_1)\sum_{i=1}^{m}(x_i - z_1) = (z_1 - \theta_1)\left(\sum_{i=1}^{m}x_i - mz_1\right) = 0$$

so that

$$\sum_{i=1}^{m}(x_i - \theta_1)^2 = mz_2 + m(z_1 - \theta_1)^2$$

Therefore the joint pdf can be written as

$$(\theta_2\sqrt{2\pi})^{-m-1} \exp\left\{-\frac{1}{2\theta_2^{\,2}}[mz_2 + m(z_1 - \theta_1)^2] - \frac{1}{2\theta_2^{\,2}}(y - \theta_1)^2\right\}$$

This shows that z_1, z_2 are sufficient for this decision problem. z_1, z_2 are known, respectively, as the "sample mean" and "sample variance" of the "sample" consisting of the numbers $x_1, \ldots, x_m$. Since much of conventional statistical theory assumes normal distributions, our discussion illustrates why the sample mean and the sample variance have such an important role in textbooks on statistical theory.

5.14. Selecting One Particular Decision Rule. Most of our discussion so far has been devoted to methods for finding all the admissible decision rules. But since in most problems there are infinitely many admissible decision rules, what further principles can be used to select one particular decision rule from among all the admissible decision rules?

As a matter of fact, we could claim that it is not the statistician's job to select one particular decision rule, but to find all the admissible decision rules. Then the person who will actually incur the loss should select one particular decision rule from among the admissible decision rules presented to him by the statistician. Some general principles for selecting one particular decision rule have been suggested (but none has been universally adopted), and we shall describe two of these principles.

One principle uses so-called "subjective probabilities." Subjective probabilities are probabilities assigned to the possible values of θ and are supposed to represent the degree of belief that a given θ represents the true joint distribution. Thus, if there were six possible joint distributions, so that θ ran from 1 to 6, and if it were felt that the first distribution was twice as "likely" to be the true distribution as any of the other distributions, the subjective probability assigned to $\theta = 1$ would be $\frac{2}{7}$ and the subjective probability assigned to each of the other five possible values of θ would be $\frac{1}{7}$. The principle states: Choose a decision rule which is Bayes relative to $b(1), \ldots, b(h)$, where $b(\theta)$ is equal to the subjective probability assigned to the value θ. The rationale for this principle is that if the $b(\theta)$ were real probabilities instead of subjective probabilities, we should want to choose a decision rule s that minimizes $\sum_{\theta=1}^{h} b(\theta)r(\theta;s)$, since this sum would represent the expected loss if the true θ were chosen by a random device that assigns probabilities $b(1), \ldots, b(h)$ to the possible values of θ. A difficulty in applying this principle is that there are no objective methods given for assigning the subjective probabilities.

A different principle that has been suggested is the "minimax criterion." For any decision rule s, let $M(s)$ denote $\max_{\theta} r(\theta;s)$. Thus, in Example 1 of Sec. 5.6, we introduced a decision rule s_1 and found that $r(\theta;s_1)$ is given by $500 + 1{,}500\theta - 1{,}000\theta^2$ for $0 \leq \theta \leq 1$. By differentiation, we find that $r(\theta;s_1)$ is a maximum when $\theta = \frac{3}{4}$, and $r(\frac{3}{4};s_1) = (\frac{1}{2})(2{,}125)$. Therefore $M(s_1) = (\frac{1}{2})(2{,}125)$. Similarly, for the decision rule s_2 introduced in Example 1 of Sec. 5.6, we find $M(s_2) = 1{,}500$. The minimax criterion states: Use a decision rule s which minimizes $M(s)$. Such a rule is called a "minimax decision rule." For example, in Example 1 of Sec. 5.6 we saw that $r(\frac{1}{2};s) = 1{,}000$ for *any* decision rule s. But then for any decision rule s, $M(s) \geq 1{,}000$. Let t denote the decision rule with $t(2;x) = 1$ for all x; that is, t chooses decision 2 no matter what the observations are. Then $r(\theta;t) = 1{,}000$ for all θ, and $M(t) = 1{,}000$. Thus t is a minimax decision rule for Example 1 of Sec. 5.6.

The minimax criterion has been criticized as being too conservative. Thus suppose in a hypothetical problem with five possible joint distributions, we are comparing two decision rules s_1 and s_2, and the expected losses when using each are given by the following table:

	θ				
	1	2	3	4	5
$r(\theta;s_1)$	1	8	6	3	1
$r(\theta;s_2)$	0	8.01	2	1	0

$M(s_1) = 8$, $M(s_2) = 8.01$. The minimax criterion would tell us to use s_1 rather than s_2. But unless we were practically certain that the true θ is equal to 2, s_2 would seem to be a more reasonable decision rule to use, since $r(\theta;s_2) < r(\theta;s_1)$ for all θ except $\theta = 2$, and even when $\theta = 2$, $r(\theta;s_2)$ is only slightly greater than $r(\theta;s_1)$. However, this is an artificial and extreme example, and in practical applications such situations rarely occur.

In the next chapter, we discuss a method for the direct computation of minimax decision rules. We end this chapter by describing a method (for later use) of recognizing a decision rule as minimax. First let us assume that we are dealing with a decision problem with a finite number h of possible joint distributions. We have the following theorem: *If s is a Bayes decision rule relative to $b(1), \ldots, b(h)$, and if $r(\theta;s) = M(s)$ for each and every θ for which $b(\theta) > 0$, then s is a minimax decision rule.*

Proof. Suppose s were not a minimax decision rule. Then there would be a decision rule t with $M(t) < M(s)$. But then

$$\sum_{\theta=1}^{h} b(\theta)r(\theta;t) \leq M(t) < M(s) = \sum_{\theta=1}^{h} b(\theta)r(\theta;s)$$

which would imply that s is not a Bayes decision rule relative to $b(1), \ldots, b(h)$, a contradiction. This contradiction proves the theorem.

To generalize the preceding theorem to the case of an infinite number of possible distributions, we first define the phrase "$\bar{\theta}$ is a point of increase of the a priori distribution $B(\theta)$." This phrase means that either $B(\theta)$ assigns a positive probability to the point $\bar{\theta}$, or else $B(\theta)$ has a derivative $b(\bar{\theta})$ at $\bar{\theta}$ and $b(\bar{\theta}) > 0$. Then we have the following theorem: *If s is a Bayes decision rule relative to $B(\theta)$, and if $r(\bar{\theta};s) = M(s)$ for every $\bar{\theta}$ which is a point of increase of $B(\theta)$, then s is a minimax decision rule.*

Proof. Suppose s were not a minimax decision rule. Then there would be a decision rule t with $M(t) < M(s)$. Using the notation $R(s;B(\theta))$ introduced in Sec. 5.10, we see that in computing $R(s;B(\theta))$, the only points θ that matter are the points of increase of $B(\theta)$, since the other points are assigned zero probability by $B(\theta)$. Therefore $R(s;B(\theta)) = M(s)$, since $r(\bar{\theta};s) = M(s)$ if $\bar{\theta}$ is a point of increase of $B(\theta)$. Clearly, $R(t;B(\theta)) \leq M(t)$, and therefore $R(t;B(\theta)) < R(s;B(\theta))$. But this implies that s is not a Bayes decision rule relative to $B(\theta)$, a contradiction which proves the theorem.

Chapter 6

LINEAR PROGRAMMING AS A COMPUTATIONAL TOOL

6.1. Introduction. We introduced the concept of a minimax decision rule in Chap. 5 and gave theorems which enable us to recognize a decision rule as minimax, under certain circumstances. However, we still have no direct way of constructing a minimax decision rule. In the present chapter, we shall describe a method for the direct construction of minimax decision rules.

We start by discussing a nonstatistical problem of a type known as a "linear programming problem." Suppose there are three fuel types, with weight, energy content, and cost given by the following table:

	Fuel type		
	1	2	3
Weight per unit volume	15	10	20
Energy per unit volume	30	25	45
Cost per unit volume	1	0.8	1.4

These fuel types can be mixed in any proportions, and the weight and energy content of the mixture are the sum of the weights and energies of the component fuel types. It is desired to obtain a mixture of 10 units of volume whose total weight is no more than 160 units and total energy is at least 320 units, at minimum possible cost. To put this problem in symbolic form, let x_1 denote the volume of fuel type 1 that will be used in the mixture, x_2 the volume of fuel type 2, and x_3 the volume of fuel type 3. Then we must find the values of x_1, x_2, x_3 which minimize $x_1 + 0.8x_2 + 1.4x_3$, subject to the restrictions $x_1 \geq 0$, $x_2 \geq 0$, $x_3 \geq 0$, and

$$x_1 + x_2 + x_3 = 10$$
$$15x_1 + 10x_2 + 20x_3 \leq 160$$
$$30x_1 + 25x_2 + 45x_3 \geq 320$$

By introducing nonnegative "slack" variables x_4 and x_5, we can turn the last two inequalities into equalities:

$$15x_1 + 10x_2 + 20x_3 + x_4 = 160$$
$$30x_1 + 25x_2 + 45x_3 - x_5 = 320$$

In the general linear programming problem, we have m linear inequalities and/or equalities on our unknowns. Each inequality can be turned into an equality by introducing a "slack" variable into the inequality. Assuming this is done, we then have m linear equations in a certain number, say, n, of unknowns $x_1, \ldots, x_n$:

$$\begin{array}{c}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\end{array}$$

The problem is to find nonnegative values for $x_1, \ldots, x_n$ which satisfy the m equations and which minimize a given linear "objective function" $c_1x_1 + c_2x_2 + \cdots + c_nx_n$.

We denote by G the set of all n-dimensional points $(x_1, \ldots, x_n)$ such that $x_1 \geqslant 0, x_2 \geqslant 0, \ldots, x_n \geqslant 0$, and the values $x_1, x_2, \ldots, x_n$ satisfy the m given equations. We assume that G contains at least one point and that there is a finite value H such that every coordinate of every point in G is smaller than H. Both of these assumptions will hold in all actual problems we shall encounter. Our linear programming problem is to find a point in G at which the objective function is minimized. The following theorem will be useful.

Theorem. *If $(y_1, \ldots, y_n)$ and $(z_1, \ldots, z_n)$ are two different points in G with $y_j = 0$ if $z_j = 0$ and $z_j = 0$ if $y_j = 0$ for all j, there is a point $(w_1, \ldots, w_n)$ of G with more zero coordinates than $(y_1, \ldots, y_n)$ or $(z_1, \ldots, z_n)$ and with $c_1w_1 + \cdots + c_nw_n \leqslant c_1y_1 + \cdots + c_ny_n$ and $c_1w_1 + \cdots + c_nw_n \leqslant c_1z_1 + \cdots + c_nz_n$.*

Proof. Suppose that exactly r of the y's are positive, and for the sake of definiteness, suppose that $y_1, \ldots, y_r$ are positive, and therefore $y_{r+1} = \cdots = y_n = 0$. Then of course $z_1, \ldots, z_r$ are all positive, and $z_{r+1} = \cdots = z_n = 0$. r must be at least 1, since if r were 0 $(y_1, \ldots, y_n)$ and $(z_1, \ldots, z_n)$ would be the same point. The n-dimensional point

$$(\lambda y_1 + (1 - \lambda)z_1, \ldots, \lambda y_r + (1 - \lambda)z_r, 0, \ldots, 0)$$

is in G for any λ for which $\lambda y_i + (1 - \lambda)z_i > 0$ for $i = 1, \ldots, r$, since it is easily verified that the coordinates of this point satisfy our m equalities. But $\lambda y_i + (1 - \lambda)z_i > 0$ means that $\lambda > -z_i/(y_i - z_i)$ if $y_i - z_i > 0$ and $\lambda < -z_i/(y_i - z_i)$ if $y_i - z_i < 0$. Let A denote max $[-z_i/(y_i - z_i)]$ taken over all $i \leq r$ for which $y_i - z_i > 0$, and let B denote min $[-z_i/(y_i - z_i)]$ taken over all $i \leq r$ for which $y_i - z_i < 0$. This means that for any λ between A and B, the point $(\lambda y_1 + (1 - \lambda)z_1, \ldots, \lambda y_r + (1 - \lambda)z_r, 0, \ldots, 0)$ is in G. Clearly, $A < 0$, and since if $y_i - z_i < 0$, we have

$$\frac{-z_i}{y_i - z_i} = \frac{1}{1 - y_i/z_i} > 1, \quad \text{therefore} \quad B > 1$$

Also, A and B are finite numbers, because otherwise we could push λ far enough out to give a coordinate above H, violating our assumption. When λ is set equal to either A or B, at least one of the r quantities $\lambda y_1 + (1 - \lambda)z_1, \ldots, \lambda y_r + (1 - \lambda)z_r$ is equal to zero. There are three possible cases:

1. $c_1y_1 + \cdots + c_ry_r = c_1z_1 + \cdots + c_rz_r$. In this case, we take as our point $(w_1, \ldots, w_n)$ the point $(Ay_1 + (1 - A)z_1, \ldots, Ay_r + (1 - A)z_r, 0, \ldots, 0)$, and then we find that the value of the objective function at $(w_1, \ldots, w_n)$ is $A(c_1y_1 + \cdots + c_ry_r) + (1 - A)(c_1z_1 + \cdots + c_rz_r) = c_1y_1 + \cdots + c_ry_r = c_1z_1 + \cdots + c_rz_r$, and thus the point $(w_1, \ldots, w_n)$ satisfies the conclusion of our theorem.

2. $c_1y_1 + \cdots + c_ry_r < c_1z_1 + \cdots + c_rz_r$. In this case, we take as our point $(w_1, \ldots, w_n)$ the point $(By_1 + (1 - B)z_1, \ldots, By_r + (1 - B)z_r, 0, \ldots, 0)$, and then the value of the objective function at $(w_1, \ldots, w_n)$ is $c_1y_1 + \cdots + c_ry_r + (1 - B)(c_1z_1 + \cdots + c_rz_r - c_1y_1 - \cdots - c_ry_r)$, which is less than $c_1y_1 + \cdots + c_ry_r$. Thus the point $(w_1, \ldots, w_n)$ satisfies the conclusion of our theorem.

3. $c_1y_1 + \cdots + c_ry_r > c_1z_1 + \cdots + c_rz_r$. In this case, we take as our point $(w_1, \ldots, w_n)$ the point $(Ay_1 + (1 - A)z_1, \ldots, Ay_r + (1 - A)z_r, 0, \ldots, 0)$, and then the value of the objective function at $(w_1, \ldots, w_n)$ is $c_1z_1 + \cdots + c_rz_r + A(c_1y_1 + \cdots + c_ry_r - c_1z_1 - \cdots - c_rz_r)$, which is less than $c_1z_1 + \cdots + c_rz_r$. Thus the point $(w_1, \ldots, w_n)$ satisfies the conclusion of our theorem.

A point $(q_1, \ldots, q_n)$ in G will be said to have the "property U" if no other point of G has zero coordinates in exactly the same locations as the zero coordinates of $(q_1, \ldots, q_n)$. The theorem just proved shows that there is at least one point with the property U at which the objective function is minimized. For suppose that $(y_1, \ldots, y_n)$ is a point in G at which the objective function is minimized. If $(y_1, \ldots, y_n)$ does not have the property U, there is a different point $(z_1, \ldots, z_n)$ in G with zero coordinates in exactly the same places as the zero coordinates of $(y_1, \ldots, y_n)$.

Then the theorem tells us there is a point $(w_1, \ldots, w_n)$ in G at which the objective function is minimized, where $(w_1, \ldots, w_n)$ has more zero coordinates than $(y_1, \ldots, y_n)$. If $(w_1, \ldots, w_n)$ does not have the property U, we repeat the process, always getting points with more and more zero coordinates at which the objective function is minimized. This process must terminate, since there are only n coordinates, and at the termination we have a point of G with the property U at which the objective function is minimized.

Next we show that only a finite number of points of G have the property U. For if we specify the locations in which we want zero coordinates, either exactly one point in G has zero coordinates in the specified places, or else no points or more than one point in G has zero coordinates in the specified locations. Only in the first case does our specification lead to a point with the property U, and then it leads to exactly one such point. Since there are 2^n different ways of specifying the locations in which we want zero coordinates, there are at most 2^n points with the property U, and in most problems, there are far fewer than 2^n.

No point in G with fewer than $n - m$ zero coordinates can have the property U. For suppose that a point $(y_1, \ldots, y_n)$ with fewer than $n - m$ zero coordinates had the property U. Let r denote the number of positive coordinates of $(y_1, \ldots, y_n)$. Then r is greater than m. Without loss of generality, we can assume that $y_1 > 0, \ldots, y_r > 0$, $y_{r+1} = \cdots = y_n = 0$. Then

$$\begin{array}{c}
a_{11}y_1 + a_{12}y_2 + \cdots + a_{1r}y_r = b_1 \\
a_{21}y_1 + a_{22}y_2 + \cdots + a_{2r}y_r = b_2 \\
\vdots \\
a_{m1}y_1 + a_{m2}y_2 + \cdots + a_{mr}y_r = b_m
\end{array}$$

Since $r > m$, there must exist quantities $(q_1, \ldots, q_r)$ with $q_i \neq y_i$ for at least one $i \leqslant r$ and with

$$\begin{array}{c}
a_{11}q_1 + a_{12}q_2 + \cdots + a_{1r}q_r = b_1 \\
a_{21}q_1 + a_{22}q_2 + \cdots + a_{2r}q_r = b_2 \\
\vdots \\
a_{m1}q_1 + a_{m2}q_2 + \cdots + a_{mr}q_r = b_m
\end{array}$$

since otherwise m linear equations would uniquely determine r unknowns, an impossibility if $r > m$. But then a nonzero value for λ can be found so that the point $(\lambda q_1 + (1 - \lambda)y_1, \ldots, \lambda q_r + (1 - \lambda)y_r, 0, \ldots, 0)$ is in G, and $\lambda q_i + (1 - \lambda)y_i > 0$ for $i = 1, \ldots, r$, which shows that the point $(y_1, \ldots, y_n)$ does not have the property U.

Since only a finite number of points of G have the property U, and since the minimum value of the objective function is achieved at a point with the property U, we can in principle find all the points of G with the property U, compute the value of the objective function at each of these points, and use the point giving the smallest value of the objective function. As an example, we discuss the fuel-mixing example with which we introduced this chapter. In that problem, $n = 5$, $m = 3$, so we have only to examine points with at least two zero coordinates. We get the following list: $x_1 = 0$, $x_2 = 0$. Then $x_3 = 10$, $x_4 = -40$, point not in G. $x_1 = 0$, $x_3 = 0$. Then $x_2 = 10$, $x_4 = 60$, $x_5 = -70$, point not in G. $x_1 = 0$, $x_4 = 0$. Then $x_2 = 4$, $x_3 = 6$, $x_5 = 50$. This point has the property U. The objective function has the value 11.6. $x_1 = 0$, $x_5 = 0$. Then $x_2 = 6.5$, $x_3 = 3.5$, $x_4 = 25$. This point has the property U. The objective function has the value 10.1. $x_2 = 0$, $x_3 = 0$. Then $x_1 = 10$, $x_4 = 10$, $x_5 = -20$, point not in G. $x_2 = 0$, $x_4 = 0$. Then $x_1 = 8$, $x_3 = 2$, $x_5 = 10$. This point has the property U. The objective function has the value 10.8. $x_2 = 0$, $x_5 = 0$. Then $x_1 = 26/3$, $x_3 = 4/3$, $x_4 = 10/3$. This point has the property U. The objective function has the value $316/30$. $x_3 = 0$, $x_4 = 0$. Then $x_1 = 12$, $x_2 = -2$, point not in G. $x_3 = 0$, $x_5 = 0$. Then $x_1 = 14$, $x_2 = -4$, point not in G. $x_4 = 0$, $x_5 = 0$. Then $x_1 = 10$, $x_2 = -1$, point not in G. From this list, we see that as soon as we set any two coordinates equal to zero, we get either a point not in G or a point with the property U. Therefore no point with more than two coordinates equal to zero will have the property U. There are four points of G with the property U, and we see that the value of the objective function is minimized when $x_1 = 0$, $x_2 = 6.5$, $x_3 = 3.5$. This is the least expensive fuel mixture that meets our volume, weight, and energy specifications.

It is easily seen that handling more complicated problems by the complete enumeration of the points with the property U, as in the preceding example, would be prohibitively long. In the next section we describe a more efficient method for finding a point at which the objective function is minimized.

6.2. The Simplex Method for Solving Linear Programming Problems. In all the problems we shall discuss, n will be greater than m, and from now on we assume that this is so. Then any point in G with the property U has at least $n - m$ zero coordinates. In some problems, every point in G with the property U has exactly $n - m$ zero coordinates: such problems are especially simple to solve and are called "nondegenerate" problems. Until further notice, we assume that we are dealing with a nondegenerate problem, so that every point in G with the property U has exactly m positive coordinates and $n - m$ zero coordinates.

Suppose that $(g_1, \ldots, g_n)$ is a point in G with the property U. With no loss of generality, we can assume that $g_1, \ldots, g_m$ are positive and $g_{m+1} = \cdots = g_n = 0$. Then the m equations in the m unknowns $x_1, \ldots, x_m$,

$$\begin{matrix} a_{11}x_1 + \cdots + a_{1m}x_m = b_1 \\ a_{21}x_1 + \cdots + a_{2m}x_m = b_2 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mm}x_m = b_m \end{matrix}$$

have a unique solution: $x_1 = g_1, \ldots, x_m = g_m$, for if there is more than one solution, it is not difficult to show that the point $(g_1, \ldots, g_n)$ does not have the property U. The fact that there is a unique solution implies that the determinant

$$\begin{vmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{vmatrix}$$

is not equal to zero. This in turn means that we can solve the equations

$$\begin{matrix} a_{11}x_1 + \cdots + a_{1m}x_m = b_1 - a_{1,m+1}x_{m+1} - \cdots - a_{1n}x_n \\ a_{21}x_1 + \cdots + a_{2m}x_m = b_2 - a_{2,m+1}x_{m+1} - \cdots - a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mm}x_m = b_m - a_{m,m+1}x_{m+1} - \cdots - a_{mn}x_n \end{matrix}$$

for the quantities $x_1, \ldots, x_m$ in terms of the quantities $x_{m+1}, \ldots, x_n$. Each of the quantities $x_1, \ldots, x_m$ will be a linear function of the quantities $x_{m+1}, \ldots, x_n$, say,

$$x_i = g_i + h_{i,m+1}x_{m+1} + \cdots + h_{in}x_n$$

for $i = 1, \ldots, m$. (Note that when $x_{m+1} = \cdots = x_n = 0$, x_i must equal g_i, to be consistent with our original assumptions.) The objective function $c_1x_1 + \cdots + c_nx_n$ can now be expressed in terms of $x_{m+1}, \ldots, x_n$, say, as $g_0 + d_{m+1}x_{m+1} + \cdots + d_nx_n$. It is convenient to display all these relationships in the form of a "tableau," as follows:

		Coefficient of:			
	Constant	x_{m+1}	x_{m+2}	$\cdots$	x_n
x_1	g_1	$h_{1,m+1}$	$h_{1,m+2}$	$\cdots$	h_{1n}
x_2	g_2	$h_{2,m+1}$	$h_{2,m+2}$	$\cdots$	h_{2n}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
x_m	g_m	$h_{m,m+1}$	$h_{m,m+2}$	$\cdots$	h_{mn}
Objective	g_0	d_{m+1}	d_{m+2}	$\cdots$	d_n

It should be emphasized that we chose to express everything in terms of $x_{m+1}, \ldots, x_n$ purely for convenience in writing: in any specific example, different sets of x's will be displayed along the sides and top of the tableau. As a numerical example, the tableau corresponding to the point (0,4,6,0,50) of the fuel-mixing example of Sec. 6.1 is easily found to be

		x_1	x_4
x_2	4	−0.5	0.1
x_3	6	−0.5	−0.1
x_5	50	−5	−2
Objective	11.6	−0.1	−0.06

Returning now to our general tableau corresponding to the point $(g_1,g_2, \ldots, g_m, 0, \ldots, 0)$, we see that if the values $d_{m+1}, \ldots, d_n$ are all nonnegative, the objective function cannot be decreased by raising any of the quantities $x_{m+1}, \ldots, x_n$ above zero. Thus, if the values $d_{m+1}, \ldots, d_n$ are all nonnegative, the objective function is minimized at the point $(g_1,g_2, \ldots, g_m, 0, \ldots, 0)$. However, if one or more of the values $d_{m+1}, \ldots, d_n$ are negative, the objective function can be decreased by raising any of the quantities $x_{m+1}, \ldots, x_n$ corresponding to a negative d. The simplex method raises *one* of the quantities $x_{m+1}, \ldots, x_n$ corresponding to a negative d as much as possible. If precisely one of the quantities $d_{m+1}, \ldots, d_n$ is negative, there is no choice about which x is to be raised above zero. If more than one of the quantities $d_{m+1}, \ldots, d_n$ is negative, there is a choice as to which x is to be raised above zero. As a rule of thumb, we raise the x corresponding to the negative d which is largest in absolute value. In the numerical example of the preceding paragraph, x_1 would be raised above zero.

In our general tableau, suppose d_s is the largest negative d in absolute value, so that we raise x_s above zero. x_s is to be raised as much as possible: suppose x_s can be raised to Δ and no further. By the assumption made above, $\Delta < H < \infty$. When x_s is raised to Δ, while the rest of the quantities $x_{m+1}, \ldots, x_n$ are held at zero, x_i becomes $g_i + h_{is}\Delta$ for $i = 1, \ldots, m$. We must have $g_i + h_{is}\Delta \geq 0$ for $i = 1, \ldots, m$. This is a restriction on Δ only if $h_{is} < 0$, in which case $\Delta \leq (-g_i/h_{is})$. From this, it follows that $\Delta = \min\,(-g_i/h_{is})$, where the minimum is taken over all values of i between 1 and m for which h_{is} is negative. In our numerical example $\Delta = \min\,(-4/-0.5,\ -6/-0.5,\ -50/-5) = 8$. When x_s is raised to Δ, at least one of the quantities $x_1, \ldots, x_m$ becomes zero, and by our assumption of nondegeneracy, *exactly* one of the quantities $x_1, \ldots, x_m$ becomes zero. Suppose x_r becomes zero, so that $\Delta = -g_r/h_{rs}$. (In our numerical example, x_2 becomes zero.) Now we construct a new tableau, with x_r shifted to the top and x_s shifted to the left column. We distinguish the entries in this new tableau by primes: g_i', h_{ij}', d_j'. In order to develop formulas for g_i', h_{ij}', d_j' in terms of g_i, h_{ij}, d_j, we first express x_s in terms of $x_r, x_{m+1}, x_{m+2}, \ldots, x_{s-1}, x_{s+1}, \ldots, x_n$, as follows. We know from the starting tableau that

$$x_r = g_r + h_{r,m+1}x_{m+1} + \cdots + h_{rs}x_s + \cdots + h_{rn}x_n$$

Since $h_{rs} < 0$, we can solve this equation for x_s, getting

$$x_s = \frac{-g_r}{h_{rs}} + \frac{1}{h_{rs}}\,x_r - \frac{h_{r,m+1}}{h_{rs}}\,x_{m+1} - \cdots - \frac{h_{rn}}{h_{rs}}\,x_n$$

Thus we have the entries in the new tableau in the row for x_s:

$$g_s' = \frac{-g_r}{h_{rs}},\ h_{sr}' = \frac{1}{h_{rs}},\ h_{sj}' = \frac{-h_{rj}}{h_{rs}} \qquad \text{for } j \neq r$$

Now we express x_i $(1 \leq i \leq m;\ i \neq r)$ in terms of $x_r, x_{m+1}, x_{m+2}, \ldots, x_{s-1}, x_{s+1}, \ldots, x_n$, by means of the following calculation. From the first tableau,

$$x_i = g_i + h_{i,m+1}x_{m+1} + \cdots + h_{is}x_s + \cdots + h_{in}x_n$$

Substituting the equation for x_s in terms of $x_r, x_{m+1}, \ldots, x_{s-1}, x_{s+1}, \ldots, x_n$ that we developed above, we find

$$x_i = g_i - \frac{g_r h_{is}}{h_{rs}} + \frac{h_{is}}{h_{rs}}\,x_r + \left(h_{i,m+1} - \frac{h_{is}h_{r,m+1}}{h_{rs}}\right)x_{m+1} + \cdots$$
$$+ \left(h_{in} - \frac{h_{is}h_{rn}}{h_{rs}}\right)x_n$$

This gives the entries in the new tableau in the row for x_i:

$$g_i' = g_i - \frac{g_r h_{is}}{h_{rs}},\ h_{ir}' = \frac{h_{is}}{h_{rs}},\ h_{ij}' = h_{ij} - \frac{h_{is}h_{rj}}{h_{rs}} \qquad \text{for } j \neq r$$

Finally, we express the objective function in terms of $x_r, x_{m+1}, x_{m+2}, \ldots, x_{s-1}, x_{s+1}, \ldots, x_n$ by means of the following calculation. The first tableau gives that the objective function equals $g_0 + d_{m+1}x_{m+1} + \cdots + d_s x_s + \cdots + d_n x_n$. Substituting the equation for x_s, we find that the objective function equals

$$g_0 - \frac{d_s g_r}{h_{rs}} + \frac{d_s}{h_{rs}} x_r + \left(d_{m+1} - \frac{d_s h_{r,m+1}}{h_{rs}}\right)x_{m+1} + \cdots + \left(d_n - \frac{d_s h_{rn}}{h_{rs}}\right)x_n$$

This gives the entries in the new tableau in the row for the objective function:

$$g_0' = g_0 - \frac{d_s g_r}{h_{rs}},\ d_r' = \frac{d_s}{h_{rs}},\ d_j' = d_j - \frac{d_s h_{rj}}{h_{rs}} \qquad \text{for } j \neq r$$

In our numerical example, x_r is x_2 and x_s is x_1, so the second tableau is

		x_2	x_4
x_1	$\frac{-4}{0.5} = 8$	$\frac{1}{-0.5} = -2$	$\frac{-0.1}{-0.5} = 0.2$
x_3	$6 - \frac{4(-0.5)}{-0.5} = 2$	$\frac{-0.5}{-0.5} = 1$	$-0.1 - \frac{(-0.5)(0.1)}{-0.5} = -0.2$
x_5	$50 - \frac{4(-5)}{-0.5} = 10$	$\frac{-5}{-0.5} = 10$	$-2 - \frac{(-5)(0.1)}{-0.5} = 3$
Objective	$11.6 - \frac{(-0.1)(4)}{-0.5} = 10.8$	$\frac{0.1}{-0.5} = 0.2$	$0.06 - \frac{(-0.1)(0.1)}{-0.5} = -0.08$

The new tableau corresponds to the following point in G: $x_i = 0$ if x_i is listed along the top of the tableau; $x_i = g_i'$ if x_i is listed along the side of the tableau. This point has the property U, since the x's listed along the top uniquely determine the x's along the side. Also, this point is an improvement over the point corresponding to the original tableau, since the objective function has been lowered in moving to the new tableau. If no d' in the new tableau is negative, the objective function is minimized at the point in G corresponding to the new tableau. But if at least one d' in the new tableau is negative, we can decrease the objective function by moving to a third tableau, using exactly the same method we employed to move from the first to the second tableau. We keep doing this until we reach a tableau which has no negative d's and thus represents a point at which the objective function is minimized. Such a tableau must be reached in a finite number of steps, since each tableau corresponds to a point with the property U and there is a finite number of such points. Thus, in the second tableau of our numerical example, d_4' is negative, so we can lower our objective function by increasing x_4. As we increase x_4, both x_3 and x_5 decrease, x_5 reaching zero first. Therefore, in the next tableau, x_5 will appear on top and x_4 will appear along the side. The next tableau is

		x_2	x_5
x_1	$8 - \frac{10(0.2)}{-3} = \frac{26}{3}$	$-2 - \frac{(10)(0.2)}{-3} \quad \frac{4}{3}$	$\frac{0.2}{-3} = -\frac{1}{15}$
x_3	$2 - \frac{10(-0.2)}{-3} = \frac{4}{3}$	$1 - \frac{(10)(-0.2)}{-3} = \frac{1}{3}$	$\frac{-0.2}{-3} \cdot \frac{1}{15}$
x_4	$\frac{-10}{-3} = \frac{10}{3}$	$\frac{-10}{-3} \quad \frac{10}{3}$	$\frac{1}{-3} = -\frac{1}{3}$
Objective	$10.8 - \frac{(-0.08)(10)}{-3} = \frac{316}{30}$	$0.2 - \frac{(-0.08)(10)}{3} \quad \frac{2}{30}$	$\frac{0.08}{-3} \quad \frac{8}{300}$

We see that the objective function can be decreased further by increasing x_2 above zero. As x_2 increases, x_1 decreases. Therefore we get as our next tableau

		x_1	x_5
x_2	6.5	−0.75	−0.05
x_3	3.5	−0.25	0.05
x_4	25	−2.5	−0.5
Objective	10.1	0.05	0.03

This tableau gives the final solution, since it is not possible to decrease the objective function by increasing either x_1 or x_5. Thus at the point (0,6.5,3.5,25,0) the objective function is minimized. This confirms what we found above, in the enumeration of all points with the property U.

We have finished our description of the main features of the simplex method, which moves from tableau to tableau by a systematic computational procedure. Each tableau represents a point in G with the property U. One question that we have not yet discussed is how the first tableau is found. In our simple numerical example it was easy to find a starting tableau, but in more elaborate problems it can be quite a task. As we shall see, whenever the simplex technique is applied to a statistical decision problem, it will be a simple matter to find a starting tableau.

For the sake of completeness, however, we sketch a method for finding a starting tableau that will work for any linear programming problem. The starting tableau is found by *first* solving the following linear programming problem: Find the nonnegative quantities $x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m}$ which minimize $x_{n+1} + \cdots + x_{n+m}$ and which satisfy the m equations:

$$\begin{array}{l}
a_{11}x_1 + \cdots + a_{1n}x_n \pm x_{n+1} = b_1 \\
a_{21}x_1 + \cdots + a_{2n}x_n \pm x_{n+2} = b_2 \\
\quad \vdots \\
a_{m1}x_1 + \cdots + a_{mn}x_n \pm x_{n+m} = b_m
\end{array}$$

where in the ith equation $+x_{n+i}$ is used if $b_i > 0$, $-x_{n+i}$ is used if $b_i < 0$. (If $b_i = 0$, x_{n+i} is *not* introduced into the problem.) For this linear programming problem, a starting tableau can be written immediately by expressing $x_{n+1}, \ldots, x_{n+m}$ in terms of $x_1, \ldots, x_n$. This tableau corresponds to the point $x_1 = \cdots = x_n = 0$, $x_{n+1} = |b_1|, \ldots, x_{n+m} = |b_m|$. The final tableau for the linear programming problem will obviously represent a point where $x_{n+1} = \cdots = x_{n+m} = 0$, since the problem was to minimize $x_{n+1} + \cdots + x_{n+m}$. But then the coordinates $x_1, \ldots, x_n$ of this point satisfy the m equalities of the *original* linear programming problem, and only m of the quantities $x_1, \ldots, x_n$ are positive. Thus the quantities $x_1, \ldots, x_n$ represent a point with the property U for our *original* linear programming problem. The tableau for the original problem corresponding to this point $x_1, \ldots, x_n$ can then be constructed.

All the computations described so far are based on the supposition that our linear programming problem is nondegenerate; that is, every point in G with the property U has exactly m positive coordinates and $n - m$ zero coordinates. What happens if this is not so? We may reach a tableau in which one or more of the g_i's are zero ($i \neq 0$). Then, even though one of the d's is negative, it may be impossible to increase the corresponding x above zero because an x on the left corresponding to a zero g may be made negative, which is not allowed. One way to handle this situation is arbitrarily to increase the zero g's by very small amounts. This of course changes the original problem to a new problem, but the new problem is nondegenerate and is very close to the original problem. Once we get a solution for the new problem, we can usually recognize the solution to the original problem: If an x in the solution to the new problem is very close to zero, the corresponding x in the solution to the original problem is zero. For details, we refer the reader to a text on linear programming.

6.3. Application of Linear Programming to Statistical Decision Problems. Suppose we want to construct a minimax decision rule s for a decision problem in which there is a finite number L of possible decisions, a finite number h of possible distributions, and each distribution allows a finite number of possible values for x. Denote by q the total number of different possible values for x allowed by the whole set of distributions. We assume that $W(y;D;x) \geq 0$ for all y, D, x. If necessary, we can add a positive constant C to $W(y;D;x)$ to make this so. The addition of C to $W(y;D;x)$ simply replaces $r(\theta;s)$ by $r(\theta;s) + C$ for all θ and does not change the minimax decision rule. Under this assumption, $M(s) \geqslant 0$.

From formulas developed in Chap. 5,

$$r(\theta;s) = \sum_D \sum_x s(D;x)\left[\sum_y W(y;D;x)f(x,y;\theta)\right]$$

We denote $\sum_y W(y;D;x)f(x,y;\theta)$ by $A(D;x;\theta)$. Thus

$$r(\theta;s) = \sum_D \sum_x A(D;x;\theta)s(D;x)$$

Define quantities $z_1, \ldots, z_h$ by the equations

$$\begin{aligned}
\sum_D \sum_x A(D;x;1)s(D;x) + z_1 &= M(s)\\
\sum_D \sum_x A(D;x;2)s(D;x) + z_2 &= M(s)\\
\vdots \qquad\qquad & \qquad \vdots\\
\sum_D \sum_x A(D;x;h)s(D;x) + z_h &= M(s)
\end{aligned}$$

Since $r(\theta;s) \leq M(s)$ for all θ, $z_1, z_2, \ldots, z_h$ are all nonnegative. $A(D;x;\theta)$ are known quantities for all D, x, and θ. We set up a linear programming problem as follows. The unknowns are $M(s)$, $z_1, \ldots, z_h$, $s(D;x)$ for all D, x. There are $1 + h + Lq$ in all. The problem is to find nonnegative values for these unknowns which satisfy the $q + h$ equations

$$s(1;x) + s(2;x) + \cdots + s(L;x) = 1 \qquad \text{(one equation for each } x\text{)}$$

$$\sum_D \sum_x A(D;x;\theta)s(D;x) + z_\theta - M(s) = 0 \qquad \text{(one equation for each } \theta\text{)}$$

and which minimize $M(s)$. The values of $s(D;x)$ that we get from the computation give a minimax decision rule s.

For the linear programming problem just described, we know that any point with the property U must have at least $1 + h + Lq - q - h = 1 + q(L - 1)$ zero coordinates. We can construct the following point with the property U:

$$s(1;x) = 1 \qquad \text{for all } x$$

$$s(2;x) = \cdots = s(L;x) = 0 \qquad \text{for all } x$$

$$M(s) = \max \left[\sum_x A(1;x;1), \sum_x A(1;x;2), \ldots, \sum_x A(1;x;h) \right]$$

$$z_\theta = M(s) - \sum_x A(1;x;\theta) \qquad \text{for } \theta = 1, \ldots, h$$

Note that $q(L - 1)$ of the quantities $s(D;x)$ have been set equal to zero, and at least one of the quantities $z_1, \ldots, z_h$ has been set equal to zero by our choice for the value of $M(s)$. Of course, there are many other simple ways to find a point with the property U. Once we have the point, the corresponding tableau is easily constructed.

As a numerical example, suppose a store which has a storage tank available that can hold only one kind of fertilizer at a time has to decide whether to fill the tank with fertilizer of type 1 or of type 2. In either case, the tank holds enough for 100 sales. There are many consumers in the region, and the customers who will actually buy the fertilizer are assumed to be 100 consumers drawn at random from all the consumers in the region. The consumers are broken into two types, and the price charged depends on the consumer type and fertilizer type. The net profit on each sale is given by the following table:

		Customer type	
		1	2
Fertilizer type	1	100	40
	2	30	120

The proportion p of consumers in the region who are of type 1 is not known exactly, but from certain census data is known to be either 0.2, 0.5, or 0.8. Before deciding which fertilizer type to stock, the store will choose two consumers at random and observe which types these two consumers are. The problem is to find a minimax decision rule.

We set up the following notation. Y denotes the number of the 100 customers who will buy the fertilizer who will be of customer type 1. From the description above, it is reasonable to assume that Y has a binomial distribution with parameters $100, p$. X_1 is defined to be 1 if the first observed consumer is of type 1; 0 if the first observed consumer is of type 2. X_2 is defined to be 1 if the second observed consumer is of type 1; 0 if the second observed consumer is of type 2. X_1, X_2, Y are all independent, and each X has the probability distribution

Possible value	0	1
Probability	$1-p$	p

We assume that there is no cost connected with observing X_1 and X_2. Then the loss does not depend on X_1 or X_2, and since

$$f(x_1,x_2,y;p) = p^{x_1+x_2}(1-p)^{2-(x_1+x_2)} \frac{100!}{y!\,(100-y)!} p^y(1-p)^{100-y}$$

we see that $X_1 + X_2$ is sufficient for this problem. Denote $X_1 + X_2$ by T. Then

$$f(t,y;p) = \frac{2!}{t!\,(2-t)!} p^t(1-p)^{2-t} \frac{100!}{y!\,(100-y)!} p^y(1-p)^{100-y}$$

where the possible values for T are 0, 1, 2. We denote the decision to stock fertilizer type 1 as decision 1 and the decision to stock fertilizer type 2 as decision 2. Then

$$W(Y;1) = -100Y - 40(100 - Y) = -60Y - 4{,}000$$

$$W(Y;2) = -30Y - 120(100 - Y) = 90Y - 12{,}000$$

To ensure that $W(Y;D)$ is never negative, we add 12,000 to each $W(Y;D)$, using for the rest of the computation

$$W(Y;1) = 8{,}000 - 60Y$$

$$W(Y;2) = 90Y$$

We set $\theta = 1$ to indicate that $p = 0.2$; $\theta = 2$ to indicate that $p = 0.5$; and $\theta = 3$ to indicate that $p = 0.8$. In computing $A(D;t;\theta)$ we note that $\sum_y W(y;1)f(t,y;p)$ is equal to

$$\frac{2!}{t!\,(2-t)!} p^t(1-p)^{2-t}\left[8{,}000 - 60\sum_y y \frac{100!}{y!\,(100-y)!} p^y(1-p)^{100-y}\right]$$

$$= \frac{2!}{t!\,(2-t)!} p^t(1-p)^{2-t}(8{,}000 - 6{,}000p)$$

and
$$\sum_y W(y;2)f(t,y;p) = \frac{2!}{t!\,(2-t)!} p^t(1-p)^{2-t}9{,}000p$$

This gives

$$\begin{aligned}
A(1;0;1) &= 4{,}352 & A(2;0;1) &= 1{,}152\\
A(1;1;1) &= 2{,}176 & A(2;1;1) &= 576\\
A(1;2;1) &= 272 & A(2;2;1) &= 72\\
A(1;0;2) &= 1{,}250 & A(2;0;2) &= 1{,}125\\
A(1;1;2) &= 2{,}500 & A(2;1;2) &= 2{,}250\\
A(1;2;2) &= 1{,}250 & A(2;2;2) &= 1{,}125\\
A(1;0;3) &= 128 & A(2;0;3) &= 288\\
A(1;1;3) &= 1{,}024 & A(2;1;3) &= 2{,}304\\
A(1;2;3) &= 2{,}048 & A(2;2;3) &= 4{,}608
\end{aligned}$$

We start our construction of a point with the property U by setting the values $s(D;t)$ as follows:

$$\begin{aligned}
s(2;0) &= 1 & s(1;0) &= 0\\
s(2;1) &= 1 & s(1;1) &= 0\\
s(1;2) &= 1 & s(2;2) &= 0
\end{aligned}$$

Using these values, we find

$$\sum_D \sum_t A(D;t;1)s(D;t) = 1{,}152 + 576 + 272 = 2{,}000$$

$$\sum_D \sum_t A(D;t;2)s(D;t) = 1{,}125 + 2{,}250 + 1{,}250 = 4{,}625$$

$$\sum_D \sum_t A(D;t;3)s(D;t) = 288 + 2{,}304 + 2{,}048 = 4{,}640$$

Thus we set $M(s) = 4{,}640$, so $z_3 = 0$, $z_2 = 15$, $z_1 = 2{,}640$. Therefore the tableau corresponding to this point expresses $s(2;0)$, $s(2;1)$, $s(1;2)$, z_1, z_2, and $M(s)$ in terms of $s(1;0)$, $s(1;1)$, $s(2;2)$, and z_3. This tableau is

		$s(1;0)$	$s(1;1)$	$s(2;2)$	z_3
$s(2;0)$	1	−1	0	0	0
$s(2;1)$	1	0	−1	0	0
$s(1;2)$	1	0	0	−1	0
z_1	2,640	−3,360	−2,880	2,760	1
z_2	15	−285	−1,530	2,685	1
$M(s)$	4,640	−160	−1,280	2,560	1

Since the objective function is $M(s)$, it is not necessary to add a separate row for the objective function. We raise $s(1;1)$ to 15/1,530, making z_2 equal to zero, and thus getting as our second tableau:

		$s(1;0)$	$s(2;2)$	z_3	z_2
$s(1;1)$	$\frac{15}{1,530}$	$\frac{-285}{1,530}$	$\frac{2,685}{1,530}$	$\frac{1}{1,530}$	$\frac{-1}{1,530}$
$s(2;0)$	1	-1	0	0	0
$s(2;1)$	$\frac{1,515}{1,530}$	$\frac{285}{1,530}$	$\frac{-2,685}{1,530}$	$\frac{-1}{1,530}$	$\frac{1}{1,530}$
$s(1;2)$	1	0	-1	0	0
z_1	$2,640 - \frac{2,880}{102}$	$-3,360 + \frac{9,120}{17}$	$\frac{-39,000}{17}$	$\frac{-135}{153}$	$\frac{2,880}{1,530}$
$M(s)$	$4,640 - \frac{1,280}{102}$	$\frac{4,000}{51}$	$\frac{16,000}{51}$	$\frac{25}{153}$	$\frac{1,280}{1,530}$

We see that no further improvement in $M(s)$ is possible. Thus a minimax decision rule s is given by $s(1;0) = 0$, $s(2;0) = 1$, $s(1;1) = 15/1,530$, $s(2;1) = 1,515/1,530$, $s(1;2) = 1$, $s(2;2) = 0$. For this decision rule s, $M(s) = 4,640 - 1,280/102 - 12,000$ (subtracting the 12,000 that we added at the beginning of our computation).

6.4. Finding Approximately Minimax Decision Rules. In the preceding section, we have seen that linear programming can be used to construct a minimax decision rule in a problem with a finite number of decisions, distributions, and possible values for x. If these conditions are not satisfied, linear programming cannot be used. However, under certain circumstances, we can find a decision rule which is approximately minimax by the use of linear programming.

For example, suppose that the possible distributions are those given as a parameter θ varies continuously between 0 and 1. Suppose we arbitrarily assume that the only possible distributions are those given by a finite number of values of θ spaced equally over the interval (0,1). If $r(\theta;s)$ is a continuous function of θ, and if our finite number of values of θ are spaced closely enough together, a minimax decision rule for the finite problem will be close to a minimax decision rule for the original problem.

As another example, suppose our decision is to be based on a chance variable X which can vary continuously over an interval from A to B. Suppose we modify the problem by breaking the interval (A,B) into k small subintervals and, whenever X falls into the ith subinterval, arbitrarily place its value at the midpoint of this subinterval, say, a_i. This changes the problem into one where the possible values of X are $a_1, a_2, \ldots, a_k$. [Under the θth distribution, $P(X = a_i)$ is equal to the probability assigned to the ith subinterval by the θth distribution of the original problem.] If the subintervals are small, a minimax decision rule for the new finite problem will be close to a minimax decision rule for the original problem.

In each of the cases discussed, the number of unknowns in the resulting linear programming problem is likely to be large. The use of modern computing equipment would be necessary.

Chapter 7

PROBLEMS INVOLVING A SEQUENCE OF DECISIONS OVER TIME

7.1. Introduction. Many important problems involve making a sequence of decisions at different times, rather than making a decision at only one time, as in the problems we have been discussing up to now. Then at a given time we must take into account the effect of the decision we choose on the whole future duration of the problem; that is, we cannot simply choose the decision which looks best for the immediate future, for such a decision may cause grave losses in the more distant future. We can compare the problem with that faced by a mountain climber confronted by a choice between two paths. One path may look steep and stony, and the other may look gradual and smooth, but the more pleasant-looking path may finally lead to a sheer cliff which is impossible to climb, while the stony path may in fact provide a reasonable way to the summit. In making any decision, we must take into account the whole future, not just the immediate future.

Some problems involve making decisions about which variables $X_1, \ldots, X_m$ to observe, or how many variables to observe. To distinguish such decisions from the sort of decision we have been discussing up to now, we shall call them "sampling decisions."

7.2. Problems Where There Is One Time When a Sampling Decision Is to Be Made. In problems where a sampling decision is to be made at one time only, the sampling decision chooses the chance variables on which the regular decision is to be based. (The sampling decision may be to observe *no* chance variables.)

Once the sampling decision has been made, so that the chance variables on which the regular decision is to be based are specified, we have reached a decision problem of the type we discussed in Chaps. 5 and 6, and all the techniques developed there may be used. A complete decision rule must specify how the sampling decision is to be made, and how the regular decision is to be made on the basis of the chance variables specified by the sampling decision. To find a minimax decision rule, we proceed as follows. For any sampling decision d, we can find a minimax decision rule s_d for the remaining part of the problem. As usual, $M(s_d)$ denotes $\max_{\theta} r(\theta;s_d)$. Then an over-all minimax decision rule is given as follows: Choose the sampling decision d that minimizes $M(s_d)$, and then use s_d for the remaining part of the problem.

As a simple example, we modify the numerical example of Sec. 6.3 so that the net profit on each sale is given by the following table:

		Customer type 1	Customer type 2
Fertilizer type	1	100	40
	2	40	100

Also, the proportion p of consumers in the region who are of type 1 is known to be either 0.2 or 0.8. Before deciding which fertilizer to stock, the store can choose any desired number m of consumers at random and observe which types these consumers are. However, it costs $100 to observe each consumer. The problem is to find a minimax decision rule.

In solving this problem, we note that the specification of m is part of the decision rule. Our first step is to find a minimax decision rule for each possible value of m. As in Sec. 6.3, we let T denote the number of consumers of type 1 among the m consumers observed. The loss function is

$$W(Y;1) = 100m - 100Y - 40(100 - Y) = 100m - 60Y - 4{,}000$$

$$W(Y;2) = 100m - 40Y - 100(100 - Y) = 100m + 60Y - 10{,}000$$

As in Sec. 6.3, it is easily verified that T is sufficient and that

$$f(t,y;p) = \frac{m!}{t!\,(m-t)!}\,p^t(1-p)^{m-t}\,\frac{100!}{y!\,(100-y)!}\,p^y(1-p)^{100-y}$$

Next we construct a Bayes decision rule relative to ½, ½. We have

$$K(1;t) = \frac{m!}{t!\,(m-t)!}\,[100m - 2{,}600(0.2)^t(0.8)^{m-t} - 4{,}400(0.8)^t(0.2)^{m-t}]$$

$$K(2;t) = \frac{m!}{t!\,(m-t)!}\,[100m - 4{,}400(0.2)^t(0.8)^{m-t} - 2{,}600(0.8)^t(0.2)^{m-t}]$$

From this it is easily found that

$$\begin{aligned} K(1;t) &< K(2;t) && \text{if } t > (\tfrac{1}{2})m \\ K(1;t) &= K(2;t) && \text{if } t = (\tfrac{1}{2})m \\ K(1;t) &> K(2;t) && \text{if } t < (\tfrac{1}{2})m \end{aligned}$$

Therefore the decision rule s_m defined as follows is a Bayes decision rule relative to ½, ½: $s_m(1;t) = 1$ if $t > (\frac{1}{2})m$, $s_m(1;t) = \frac{1}{2}$ if $t = (\frac{1}{2})m$, $s_m(1;t) = 0$ if $t < (\frac{1}{2})m$. (Of course, a value of t equal to $(\frac{1}{2})m$ could be observed only if m is an even number.) Then if m is odd, we find

$$r(0.2;s_m) = 100m - 8{,}800 \sum_{t=0}^{(m-1)/2} \frac{m!}{t!\,(m-t)!}(0.2)^t(0.8)^{m-t} - 5{,}200 \sum_{t=(m+1)/2}^{m} \frac{m!}{t!\,(m-t)!}(0.2)^t(0.8)^{m-t}$$

$$r(0.8;s_m) = 100m - 5{,}200 \sum_{t=0}^{(m-1)/2} \frac{m!}{t!\,(m-t)!}(0.8)^t(0.2)^{m-t} - 8{,}800 \sum_{t=(m+1)/2}^{m} \frac{m!}{t!\,(m-t)!}(0.8)^t(0.2)^{m-t}$$

Since

$$\sum_{t=0}^{(m-1)/2} \frac{m!}{t!\,(m-t)!}(0.2)^t(0.8)^{m-t} = \sum_{t=(m+1)/2}^{m} \frac{m!}{t!\,(m-t)!}(0.8)^t(0.2)^{m-t}$$

we have $r(0.2;s_m) = r(0.8;s_m) = M(s_m)$, and thus s_m is a minimax decision rule for the given m, by the theorem of Sec. 5.14. If m is even, we find

$$\begin{aligned} r(0.2;s_m) = 100m &- 8{,}800 \sum_{t=0}^{m/2-1} \frac{m!}{t!\,(m-t)!}(0.2)^t(0.8)^{m-t} \\ &- 5{,}200 \sum_{t=m/2+1}^{m} \frac{m!}{t!\,(m-t)!}(0.2)^t(0.8)^{m-t} \\ &- \tfrac{1}{2}(5{,}200 + 8{,}800)\frac{m!}{(m/2)!\,(m-m/2)!}(0.2)^{m/2}(0.8)^{m/2} \end{aligned}$$

$$\begin{aligned} r(0.8;s_m) = 100m &- 5{,}200 \sum_{t=0}^{m/2-1} \frac{m!}{t!\,(m-t)!}(0.8)^t(0.2)^{m-t} \\ &- 8{,}800 \sum_{t=m/2+1}^{m} \frac{m!}{t!\,(m-t)!}(0.8)^t(0.2)^{m-t} \\ &- \tfrac{1}{2}(5{,}200 + 8{,}800)\frac{m!}{(m/2)!\,(m-m/2)!}(0.8)^{m/2}(0.2)^{m/2} \end{aligned}$$

and we find as above that $r(0.2;s_m) = r(0.8;s_m)$, and thus s_m is a minimax decision rule for the given m.

We have now succeeded in finding a minimax decision rule for each specific choice of m. With the help of the formulas developed above and a table of the binomial distribution, we can find the numerical value of $M(s_m)$, for each m. These values are given in the following table:

m	$M(s_m)$
1	$-7{,}980$
2	$-7{,}880$
3	$-8{,}125.6$
4	$-8{,}025.6$
5	$-8{,}091.56$
6	$-7{,}991.38$
7	$-7{,}980.12$
.	.
.	.
.	.

From the table, it can be seen that the best choice of m is 3. Thus an over-all minimax decision rule for the problem can be described as follows: Choose three consumers at random, and stock fertilizer type 2 if there are 0 or 1 consumer of type 1 among the three consumers observed; stock fertilizer type 1 if there are 2 or 3 consumers of type 1 among the three consumers observed.

7.3. Sequential Sampling Problems. A sequential sampling problem is a problem in which a sampling decision must be made at more than one time. Many of these problems are very complicated, and much research remains to be done in this area. We shall discuss one of the simplest problems of this type, in order to illustrate some of the techniques which have been found useful.

We assume that the chance variables $X_1, X_2, \ldots$ are distributed independently of each other and of the chance variables $Y_1, \ldots, Y_n$. The distributions of $X_1, X_2, \ldots$ are all the same, and the common cdf is known to be one of the two given cdf's $F_1(x)$ or $F_2(x)$. If the common cdf of the X's is $F_1(x)$, then the joint cdf of $Y_1, \ldots, Y_n$ is $G_1(y_1, \ldots, y_n)$, while if the common cdf of the X's is $F_2(x)$, the joint cdf of $Y_1, \ldots, Y_n$ is $G_2(y_1, \ldots, y_n)$. We are allowed the utmost freedom in taking observations: we may observe no X's, or we may observe X_1 and then decide whether to observe X_2 or stop and choose a regular decision; if we decide to observe X_2, we can decide either to observe X_3 or to stop and choose a regular decision, etc. There are two possible regular decisions, which we shall label d_1, d_2. If we observe exactly m of the X's before stopping and choosing a regular decision, the loss incurred is $W(Y_1, \ldots, Y_n; D) + cm$, where c is a given positive constant and D can take the possible values d_1, d_2. Clearly, this means that it costs us an amount c to observe each X.

Let us denote the expected value of $W(Y_1, \ldots, Y_n; d_i)$ when the joint cdf of $Y_1, \ldots, Y_n$ is $G_j(y_1, \ldots, y_n)$ by $a(G_j;d_i)$. Under any one of the following conditions, no admissible decision rule will observe any X's.:

(1) $$a(G_1;d_1) \leq a(G_1;d_2) \qquad \text{and} \qquad a(G_2;d_1) \leq a(G_2;d_2)$$

In this case, d_1 is at least as good a decision as d_2 no matter which distribution is the true one, so we may as well choose d_1 without taking any observations and save ourselves the charge of c per observation.

(2) $$a(G_1;d_2) \leq a(G_1;d_1) \qquad \text{and} \qquad a(G_2;d_2) \leq a(G_2;d_1)$$

This is the same as situation 1 with the roles of d_1 and d_2 reversed.

(3) $$a(G_j;d_i) \leq c \qquad \text{for } i = 1, 2 \text{ and } j = 1, 2$$

In this case the cost per observation is so large that it is worthwhile to choose one of the decisions d_1 or d_2 without taking any observations at all. From now on we assume that none of the situations 1, 2, or 3 holds. Then it is no loss of generality to assume that the decisions are labeled so that

$$\begin{aligned} a(G_1;d_1) &\leq a(G_1;d_2) \\ a(G_2;d_2) &\leq a(G_2;d_1) \end{aligned} \tag{7.1}$$

with at least one of these a strict inequality. Briefly, this means that if G_1 is the distribution, we should prefer to choose d_1, and if G_2 is the distribution, we should prefer to choose d_2.

Let s_1 denote the decision rule that chooses d_1 without observing any X's, and let s_2 denote the decision rule that chooses d_2 without observing any X's. It is easily seen that

$$\begin{aligned} r(1;s_1) &= a(G_1;d_1) \\ r(2;s_1) &= a(G_2;d_1) \\ r(1;s_2) &= a(G_1;d_2) \\ r(2;s_2) &= a(G_2;d_2) \end{aligned} \tag{7.2}$$

We shall prove the following theorem.

Theorem 1. *There are values L_1, L_2, where $0 \leq L_1 \leq L_2 \leq 1$, such that s_1 is a Bayes decision rule relative to $b, 1 - b$ for every b with $L_2 \leq b \leq 1$; s_2 is a Bayes decision rule relative to $b, 1 - b$ for every b with $0 \leq b \leq L_1$; and if $L_1 < b < L_2$, then a Bayes decision rule relative to $b, 1 - b$ certainly observes X_1.*

Proof. First we note that s_1 is certainly a Bayes decision rule relative to 1, 0, since a Bayes decision rule relative to 1, 0 must minimize $r(1;s)$, and the smallest possible value for $r(1;s)$ is $a(G_1;d_1)$, which is achieved by choosing the "right" decision d_1 without taking any observations; in other words, by using the decision rule s_1. Similarly, it is easily seen that s_2 is a Bayes decision rule relative to 0, 1.

Now suppose that there were values b and b', with $b' < b$, such that s_1 is a Bayes decision rule relative to b', $1 - b'$ but not relative to b, $1 - b$. This implies that $b < 1$. There must then be a decision rule s such that

$$br(1;s) + (1 - b)r(2;s) < br(1;s_1) + (1 - b)r(2;s_1) \tag{7.3}$$

On the other hand, we have

$$b'r(1;s_1) + (1 - b')r(2;s_1) \leqslant b'r(1;s) + (1 - b')r(2;s) \tag{7.4}$$

We know that $r(1;s_1) \leqslant r(1;s)$, and therefore it follows from (7.3) that $r(2;s_1) > r(2;s)$. Then from (7.4) we find $r(1;s_1) < r(1;s)$. The inequality (7.3) is equivalent to

$$\frac{b}{1 - b} < \frac{r(2;s_1) - r(2;s)}{r(1;s) - r(1;s_1)} \tag{7.5}$$

and the inequality (7.4) is equivalent to

$$\frac{b'}{1 - b'} \geqslant \frac{r(2;s_1) - r(2;s)}{r(1;s) - r(1;s_1)} \tag{7.6}$$

But (7.5) and (7.6) taken together imply that $b' > b$, which contradicts our assumption that $b' < b$. This means that if s_1 is a Bayes decision rule relative to b', $1 - b'$, then s_1 is a Bayes decision rule relative to b, $1 - b$ for all $b > b'$. Let L_2 denote the smallest number with the property that s_1 is a Bayes decision rule relative to b, $1 - b$ for every $b > L_2$. Our discussion above shows that such an L_2 exists. We now show that s_1 is a Bayes decision rule relative to L_2, $1 - L_2$. For suppose it were not. Then there would be a decision rule t with

$$L_2r(1;t) + (1 - L_2)r(2;t) < L_2r(1;s_1) + (1 - L_2)r(2;s_1)$$

But then there would be a value b' slightly above L_2 such that

$$b'r(1;t) + (1 - b')r(2;t) < b'r(1;s_1) + (1 - b')r(2;s_1)$$

which would contradict the fact that s_1 is a Bayes decision rule relative to b', $1 - b'$. This contradiction proves that s_1 is a Bayes decision rule relative to L_2, $1 - L_2$. From the definition of L_2, it is clear that s_1 is not a Bayes decision rule relative to b, $1 - b$ if b is any value below L_2.

In an analogous manner, we can prove the existence of a value L_1, such that s_2 is a Bayes decision rule relative to b, $1 - b$ for every $b \leqslant L_1$, but s_2 is not a Bayes decision rule relative to b, $1 - b$ if b is any value above L_1.

Our next task is to show that $L_1 \leqslant L_2$. For suppose that $L_1 > L_2$. Then both s_1 and s_2 would be Bayes decision rules relative to b, $1 - b$ for all values b between L_2 and L_1. Then for any such b, we must have

$$br(1;s_1) + (1 - b)r(2;s_1) = br(1;s_2) + (1 - b)r(2;s_2) \tag{7.7}$$

or, using (7.2),

$$ba(G_1;d_1) + (1 - b)a(G_2;d_1) = ba(G_1;d_2) + (1 - b)a(G_2;d_2)$$

which implies that

$$b = \frac{a(G_2;d_1) - a(G_2;d_2)}{a(G_2;d_1) - a(G_2;d_2) + a(G_1;d_2) - a(G_1;d_1)}$$

From (7.1), the denominator of this fraction is positive, and thus equation (7.7) determines a unique value of b. But this contradicts the fact that (7.7) must hold for all b between L_2 and L_1 and shows that $L_1 \leqslant L_2$.

To complete the proof of Theorem 1, we must show that if $L_1 < b < L_2$, then a Bayes decision rule relative to b, $1 - b$ certainly observes X_1. To show this, let b be a fixed value with $L_1 < b < L_2$, and let s denote a Bayes decision rule relative to b, $1 - b$. Let p_1 denote the probability assigned by s to choosing d_1 without observing X_1, and let p_2 denote the probability assigned by s to choosing d_2 without observing X_1. If $p_1 + p_2 = 1$, s certainly chooses a decision without observing X_1, and clearly

$$\begin{aligned} r(1;s) &= p_1a(G_1;d_1) + p_2a(G_1;d_2) \\ r(2;s) &= p_1a(G_2;d_1) + p_2a(G_2;d_2) \end{aligned} \tag{7.8}$$

Also

$$\begin{aligned} br(1;s) + (1 - b)r(2;s) &< br(1;s_1) + (1 - b)r(2;s_1) \\ br(1;s) + (1 - b)r(2;s) &< br(1;s_2) + (1 - b)r(2;s_2) \end{aligned} \tag{7.9}$$

From (7.2) and (7.8), we find

$$\begin{aligned} br(1;s) + (1 - b)r(2;s) = p_1[br(1;s_1) &+ (1 - b)r(2;s_1)] \\ &+ p_2[br(1;s_2) + (1 - b)r(2;s_2)] \end{aligned} \tag{7.10}$$

But (7.9) and (7.10) contradict each other, and this contradiction proves that $p_1 + p_2 < 1$.

Since $p_1 + p_2 < 1$, there is a positive probability $1 - p_1 - p_2$ assigned by the decision rule s to observing X_1. Let t denote the decision rule that observes X_1 with probability 1, and thereafter makes exactly the same decisions as s does once s observes X_1. Then we have

$$\begin{aligned} r(1;s) &= p_1a(G_1;d_1) + p_2a(G_1;d_2) + (1 - p_1 - p_2)r(1;t) \\ r(2;s) &= p_1a(G_2;d_1) + p_2a(G_2;d_2) + (1 - p_1 - p_2)r(2;t) \end{aligned} \tag{7.11}$$

From (7.2) and (7.11), we find

$$\begin{aligned} br(1;s) + (1-b)r(2;s) = p_1[br(1;s_1) + (1-b)r(2;s_1)] \\ + p_2[br(1;s_2) + (1-b)r(2;s_2)] \\ + (1-p_1-p_2)[br(1;t) + (1-b)r(2;t)] \end{aligned} \tag{7.12}$$

Since $$br(1;s) + (1-b)r(2;s) < br(1;s_1) + (1-b)r(2;s_1)$$

and $$br(1;s) + (1-b)r(2;s) < br(1;s_2) + (1-b)r(2;s_2)$$

it follows from (7.12) that if $p_1 + p_2 > 0$, we would have to have

$$br(1;t) + (1-b)r(2;t) < br(1;s) + (1-b)r(2;s)$$

which contradicts the fact that s is a Bayes decision rule relative to b, $1-b$. This means that $p_1 + p_2 = 0$, so that s certainly observes X_1. This completes the proof of Theorem 1.

Before stating Theorem 2, we introduce the following notation. If $F_i(x)$ has a derivative, $f_i(x)$ denotes this derivative. If $F_i(x)$ increases only in jumps, then $f_i(x)$ denotes the $P(X = x)$ assigned by $F_i(x)$. Now we prove the following theorem.

Theorem 2. *Suppose b a given value with $L_1 < b < L_2$. Let t_b denote the decision rule described as follows. t_b observes $X_1, \ldots, X_n$, where n is the smallest positive integer for which it is not true that*

$$\frac{b}{1-b}\left(\frac{1}{L_2} - 1\right) < \frac{f_2(X_1)\cdots f_2(X_n)}{f_1(X_1)\cdots f_1(X_n)} < \frac{b}{1-b}\left(\frac{1}{L_1} - 1\right)$$

If $$\frac{f_2(X_1)\cdots f_2(X_n)}{f_1(X_1)\cdots f_1(X_n)} \leq \frac{b}{1-b}\left(\frac{1}{L_2} - 1\right) \qquad t_b \text{ chooses } d_1$$

If $$\frac{f_2(X_1)\cdots f_2(X_n)}{f_1(X_1)\cdots f_1(X_n)} \geq \frac{b}{1-b}\left(\frac{1}{L_1} - 1\right) \qquad t_b \text{ chooses } d_2$$

Then t_b is a Bayes decision rule relative to b, $1-b$.

Proof. For typographical simplicity, we denote

$$bf_1(x_1)\cdots f_1(x_m) + (1-b)f_2(x_1)\cdots f_2(x_m) \text{ by } q(x_1, \ldots, x_m)$$

for any positive integer m.

Suppose s is a Bayes decision rule relative to b, $1-b$. From Theorem 1, we know that s certainly observes X_1. For any positive integer m, let $r(i; s \mid x_1, \ldots, x_m)$ denote the conditional expected loss when using s and G_i is the true distribution, given that we have observed $X_1, \ldots, X_m$ and that $X_1 = x_1, \ldots, X_m = x_m$. Let $g(x_1, \ldots, x_{m-1})$ denote the probability assigned by s to observing at least m X's, when the first observation is equal to x_1, the second observation is equal to $x_2, \ldots$, and the $(m-1)$st observation is equal to x_{m-1}. [Unless s used randomization in making its sampling decisions, the only possible values for $g(x_1, \ldots, x_{m-1})$ would be 0 and 1.]

For the sake of definiteness, we assume for the remainder of the discussion that $F_1(x)$ and $F_2(x)$ increase only in jumps. The case where density functions exist requires extremely simple modifications. We can write, for $i = 1$ or 2,

$$r(i;s) = K_i(m) + \sum_{x_1} \cdots \sum_{x_m} g(x_1, \ldots, x_{m-1}) f_i(x_1) \cdots f_i(x_m) r(i; s \mid x_1, \ldots, x_m)$$

Here $K_i(m)$ is the conditional expected value of the loss, given that sampling terminates before X_m is observed. The second expression on the right should not be difficult to understand, since $g(x_1, \ldots, x_{m-1})$ $f_i(x_1) \cdots f_i(x_m)$ is the probability of observing $X_1 = x_1, \ldots, X_m = x_m$ when s is used and $F_i(x)$ is the true cdf for X, while $r(i; s \mid x_1, \ldots, x_m)$ is the conditional expected loss given that we have observed $X_1 = x_1, \ldots,$ $X_m = x_m$. Then we can write

$$\begin{aligned} br(1;s) + (1-b)r(2;s) &= bK_1(m) + (1-b)K_2(m) \\ &+ \sum_{x_1} \cdots \sum_{x_m} g(x_1, \ldots, x_{m-1}) q(x_1, \ldots, x_m) \\ &\times \left[\frac{bf_1(x_1) \cdots f_1(x_m)}{q(x_1, \ldots, x_m)} r(1; s \mid x_1, \ldots, x_m) \right. \\ &\left. + \frac{(1-b)f_2(x_1) \cdots f_2(x_m)}{q(x_1, \ldots, x_m)} r(2; s \mid x_1, \ldots, x_m) \right] \end{aligned}$$

Then it is clear that s will make this last expression as small as possible only if, for each set of values $x_1, \ldots, x_m$, s minimizes

$$\frac{bf_1(x_1) \cdots f_1(x_m)}{q(x_1, \ldots, x_m)} r(1; s \mid x_1, \ldots, x_m) + \frac{(1-b)f_2(x_1) \cdots f_2(x_m)}{q(x_1, \ldots, x_m)} r(2; s \mid x_1, \ldots, x_m) \tag{7.13}$$

Next we define a decision rule $s(x_1, \ldots, x_m)$ for each set of values $x_1, \ldots, x_m$ as follows. $s(x_1, \ldots, x_m)$ uses $X_1, X_2, \ldots$ in exactly the same way as s uses $X_{m+1}, X_{m+2}, \ldots$ after s has observed $X_1 = x_1, \ldots, X_m = x_m$. For example, if after observing $X_1 = x_1, \ldots, X_m = x_m$, s stops sampling and chooses d_1, then $s(x_1, \ldots, x_m)$ chooses d_1 without observing X_1. If after observing $X_1 = x_1, \ldots, X_m = x_m$, s observes X_{m+1} and chooses d_1 if $X_{m+1} < 3$, then $s(x_1, \ldots, x_m)$ observes X_1 and chooses d_1 if $X_1 < 3$.

Since all the X's are independent, and each has the same distribution, and it costs as much to observe one X as another, we have

$$\begin{aligned} r(1; s \mid x_1, \ldots, x_m) &= r(1; s(x_1, \ldots, x_m)) + cm \\ r(2; s \mid x_1, \ldots, x_m) &= r(2; s(x_1, \ldots, x_m)) + cm \end{aligned} \tag{7.14}$$

The term cm represents the amount that must be paid for observing $X_1, \ldots, X_m$, an amount that would not be included in $r(i; s(x_1, \ldots, x_m))$, since $s(x_1, \ldots, x_m)$ is a decision rule which starts sampling at X_1. Substituting (7.14) into (7.13), we see that for each set of values $x_1, \ldots, x_m$, s should minimize

$$cm + \frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} r(1; s(x_1, \ldots, x_m)) + \frac{(1-b)f_2(x_1)\cdots f_2(x_m)}{q(x_1, \ldots, x_m)} r(2; s(x_1, \ldots, x_m)) \tag{7.15}$$

s will minimize (7.15) if $s(x_1, \ldots, x_m)$ is a Bayes decision rule relative to

$$\frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)}, \qquad \frac{(1-b)f_2(x_1)\cdots f_2(x_m)}{q(x_1, \ldots, x_m)}$$

From Theorem 1, we know that this implies that $s(x_1, \ldots, x_m)$ should choose d_1 without observing X_1 if

$$\frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} \geqslant L_2$$

should choose d_2 without observing X_1 if

$$\frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} \leqslant L_1$$

and should observe X_1 if

$$L_1 < \frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} < L_2$$

Recalling how $s(x_1, \ldots, x_m)$ was defined in terms of s, we see that for s to minimize (7.13), s should choose d_1 without observing X_{m+1} if

$$\frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} \geqslant L_2$$

s should choose d_2 without observing X_{m+1} if

$$\frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} \leqslant L_1$$

and s should observe X_{m+1} if

$$L_1 < \frac{bf_1(x_1)\cdots f_1(x_m)}{q(x_1, \ldots, x_m)} < L_2$$

A simple calculation shows that this description of s is exactly equivalent to the following: s should choose d_1 without observing X_{m-1} if

$$\frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} \leq \frac{b}{1-b}\left(\frac{1}{L_2} - 1\right)$$

s should choose d_2 without observing X_{m-1} if

$$\frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} \geq \frac{b}{1-b}\left(\frac{1}{L_1} - 1\right)$$

s should observe X_{m+1} if

$$\frac{b}{1-b}\left(\frac{1}{L_2} - 1\right) < \frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} < \frac{b}{1-b}\left(\frac{1}{L_1} - 1\right)$$

But this description of s shows that s is the same decision rule as the decision rule t_b described in the statement of Theorem 2, and thus proves the theorem, since s is a Bayes decision rule relative to b, $1 - b$.

The decision rule t_b is called a "Wald sequential rule," after its discoverer.

7.4. Finding a Minimax Wald Sequential Rule. In Sec. 7.3 we described a decision rule which is a Bayes decision rule relative to b, $1 - b$. In this section we attempt to find a minimax decision rule for the decision problem described in Sec. 7.3.

First we show that if $a(G_1;d_1) \geq a(G_2;d_1)$, then s_1 is a minimax decision rule. This is so because $r(1;s_1) = a(G_1;d_1)$ and $r(2;s_1) = a(G_2;d_1)$, and therefore $M(s_1) = r(1;s_1)$. Since it is known that s_1 is a Bayes decision rule relative to 1, 0, the theorem of Sec. 5.14 tells us that s_1 is minimax. In exactly the same way, we find that if $a(G_1;d_2) \leq a(G_2;d_2)$, then s_2 is a minimax decision rule.

From now on we assume that $a(G_1;d_1) < a(G_2;d_1)$ and $a(G_1;d_2) > a(G_2;d_2)$, which is true in all problems of practical interest. If we can find a value b with $L_1 < b < L_2$ such that the Bayes decision rule t_b relative to b, $1 - b$ has $r(1;t_b) = r(2;t_b)$, then the theorem of Sec. 5.14 shows that t_b is minimax. The description of t_b given in Sec. 7.3 shows that the decision rule is described completely once we know the values $\frac{b}{1-b}\left(\frac{1}{L_2} - 1\right)$ and $\frac{b}{1-b}\left(\frac{1}{L_1} - 1\right)$. For typographic simplicity, we denote $\frac{b}{1-b}\left(\frac{1}{L_2} - 1\right)$ by B and $\frac{b}{1-b}\left(\frac{1}{L_1} - 1\right)$ by A. Our next task is to try to find the values of A and B that make $r(1;t_b) = r(2;t_b)$, so that t_b is minimax. Actually, we cannot find the exact values of A and B that do this, so we shall have to be satisfied with approximations.

We denote by $\alpha(A,B)$ the probability of choosing d_1 when the true common distribution of the X's is $F_2(x)$ and the decision rule t_b is used. We denote by $\beta(A,B)$ the probability of choosing d_2 when the true common distribution of the X's is $F_1(x)$ and the decision rule t_b is used. When the decision rule t_b is used, it is clear that the total number of X's observed before a final decision is chosen is a chance variable, which we denote by N. Also, $n_i(A,B)$ denotes the expected value of N when the true common distribution of the X's is $F_i(x)$. Then we have

$$\begin{aligned} r(1;t_b) &= [1-\beta(A,B)]a(G_1;d_1) + \beta(A,B)a(G_1;d_2) + cn_1(A,B) \\ r(2;t_b) &= \alpha(A,B)a(G_2;d_1) + [1-\alpha(A,B)]a(G_2;d_2) + cn_2(A,B) \end{aligned} \tag{7.16}$$

We shall develop approximations for $\alpha(A,B)$, $\beta(A,B)$, $n_1(A,B)$, and $n_2(A,B)$. This will enable us to find approximately the values of A and B that make $r(1;t_b) = r(2;t_b)$.

Let $S_i(m)$ denote the set of m-dimensional points $x_1, \ldots, x_m$ such that if $X_1 = x_1, \ldots, X_m = x_m$, then t_b continues sampling to X_m and chooses d_i without observing X_{m+1}. Then, when $F_i(x)$ is the true cdf for the X's,

$$P(d_j \text{ is chosen and } N = m) = \sum_{x_1,\ldots,x_m} \cdots \text{ in } \sum_{S_j(m)} f_i(x_1)\cdots f_i(x_m)$$

At every point $x_1, \ldots, x_m$ in $S_1(m)$,

$$\frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} \leqslant B$$

and at every point $x_1, \ldots, x_m$ in $S_2(m)$,

$$\frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} \geqslant A$$

Our first approximation is to assume that at every point in $S_1(m)$,

$$\frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} = B$$

and at every point in $S_2(m)$,

$$\frac{f_2(x_1)\cdots f_2(x_m)}{f_1(x_1)\cdots f_1(x_m)} = A$$

This approximation looks drastic, but it works fairly well if the functions $f_1(x)$ and $f_2(x)$ do not differ greatly from each other, for then the ratio

$f_2(x_m)/f_1(x_m)$ will probably not move the whole ratio very far below B or above A. Using this approximation, we have

$$\begin{aligned} \sum_{S_1(m)}\cdots\sum f_2(x_1)\cdots f_2(x_m) &= B\sum_{S_1(m)}\cdots\sum f_1(x_1)\cdots f_1(x_m) \\ \sum_{S_2(m)}\cdots\sum f_2(x_1)\cdots f_2(x_m) &= A\sum_{S_2(m)}\cdots\sum f_1(x_1)\cdots f_1(x_m) \end{aligned} \tag{7.17}$$

Also

$$\begin{aligned} \alpha(A,B) &= \sum_{m=1}^{\infty}\sum_{S_1(m)}\cdots\sum f_2(x_1)\cdots f_2(x_m) \\ 1-\alpha(A,B) &= \sum_{m=1}^{\infty}\sum_{S_2(m)}\cdots\sum f_2(x_1)\cdots f_2(x_m) \\ \beta(A,B) &= \sum_{m=1}^{\infty}\sum_{S_2(m)}\cdots\sum f_1(x_1)\cdots f_1(x_m) \\ 1-\beta(A,B) &= \sum_{m=1}^{\infty}\sum_{S_1(m)}\cdots\sum f_1(x_1)\cdots f_1(x_m) \end{aligned} \tag{7.18}$$

Summing the expressions given in (7.17) with respect to m and using (7.18), we find

$$\begin{aligned} \alpha(A,B) &= B[1-\beta(A,B)] \\ 1-\alpha(A,B) &= A\beta(A,B) \end{aligned}$$

which is equivalent to

$$\begin{aligned} \beta(A,B) &= \frac{1-B}{A-B} \\ \alpha(A,B) &= B\left(\frac{A-1}{A-B}\right) \end{aligned} \tag{7.19}$$

It should be remembered that these expressions are not exact, but are approximations.

Next we develop approximations for $n_1(A,B)$ and $n_2(A,B)$. These two quantities are finite, otherwise at least one of the quantities $r(1;t_b)$, $r(2;t_b)$ would be infinite [Eq. (7.16)] and t_b would not be a Bayes decision rule relative to b, $1-b$. This means that under either $F_1(x)$ or $F_2(x)$, $\sum_{j=1}^{\infty} jP(N=j) < \infty$, so that $\sum_{j=m}^{\infty} jP(N=j)$ approaches zero as m increases. Since

$$\sum_{j=m}^{\infty} jP(N=j) \geqslant m\sum_{j=m}^{\infty} P(N=j) = mP(N \geqslant m) = m[1-P(N<m)]$$

we have shown that under either $F_1(x)$ or $F_2(x)$,

$$\lim_{m\to\infty} m[1-P(N<m)] = 0$$

Denote $\log [f_2(X_i)/f_1(X_i)]$ by Z_i. We show that under either $F_1(x)$ or $F_2(x)$, $E\{Z_1 + \cdots + Z_N\} = E\{N\}\, E\{Z_1\}$, by the following argument. For any fixed positive integer m,

$$m\, E\{Z_1\} = E\{Z_1 + \cdots + Z_m\} = \sum_{j=1}^{m} P(N = j)\, E\{Z_1 + \cdots + Z_m \mid N = j\}$$
$$+ P(N > m)\, E\{Z_1 + \cdots + Z_m \mid N > m\}$$

But $E\{Z_1 + \cdots + Z_m \mid N = j\} = E\{Z_1 + \cdots + Z_j \mid N = j\} + E\{Z_{j+1} + \cdots + Z_m \mid N = j\}$, and knowing that $N = j$ gives no information at all about $Z_{j+1}, \ldots, Z_m$, so that $E\{Z_{j+1} + \cdots + Z_m \mid N = j\} = (m - j)E\{Z_1\}$, since the marginal unconditional distributions of $Z_1, \ldots, Z_m$ are all identical. Then we have

$$m\, E\{Z_1\} = \sum_{j=1}^{m} P(N = j)\, E\{Z_1 + \cdots + Z_j \mid N = j\}$$
$$+ \sum_{j=1}^{m} P(N = j)(m - j)\, E\{Z_1\} + P(N > m)\, E\{Z_1 + \cdots + Z_m \mid N > m\} \tag{7.20}$$

Rearranging (7.20), we get

$$\sum_{j=1}^{m} P(N = j)\, E\{Z_1 + \cdots + Z_j \mid N = j\} = E\{Z_1\}[m - mP(N \leqslant m)$$
$$+ \sum_{j=1}^{m} jP(N = j)] - P(N > m)\, E\{Z_1 + \cdots + Z_m \mid N > m\} \tag{7.21}$$

Now we let m increase in (7.21). The left-hand side of (7.21) approaches $E\{Z_1 + \cdots + Z_N\}$ as m approaches infinity. The expression $m - mP(N \leqslant m)$ can be written $m[1 - P(N \leqslant m)]$, which is not larger than $m[1 - P(N < m)]$, and this last expression we know approaches zero as m approaches infinity, showing that $m - mP(N \leqslant m)$ approaches zero as m approaches infinity. $\sum_{j=1}^{m} jP(N = j)$ approaches $E\{N\}$ as m approaches infinity. Since if $N > m$ it implies that $\log B < Z_1 + \ldots + Z_m < \log A$, because of the structure of t_b, it is clear that $|E\{Z_1 + \ldots + Z_m \mid N > m\}| < \max (|\log B|, |\log A|)$. Since $P(N > m)$ approaches zero as m approaches infinity, it follows that $P(N > m)\, E\{Z_1 + \ldots + Z_m \mid N > m\}$ approaches zero as m approaches infinity. Applying these facts to (7.21) and letting m increase, we get

$$E\{Z_1 + \cdots + Z_N\} = E\{N\}\, E\{Z_1\} \tag{7.22}$$

We develop an approximation for $E\{Z_1 + \cdots + Z_N\}$ as follows. From the structure of t_b, $Z_1 + \cdots + Z_N$ is either no greater than $\log B$ (then d_1 is chosen) or no less than $\log A$ (then d_2 is chosen). Clearly, if $F_1(x)$ is the true common cdf, $P(Z_1 + \cdots + Z_N \leqslant \log B) = 1 - \beta(A,B)$, while

$P(Z_1 + \cdots + Z_N \geq \log A) = \beta(A,B)$. Similarly, if $F_2(x)$ is the true common cdf, $P(Z_1 + \cdots + Z_N \leq \log B) = \alpha(A,B)$, while $P(Z_1 + \ldots + Z_N \geq \log A) = 1 - \alpha(A,B)$. As an approximation, we compute $E\{Z_1 + \cdots + Z_N\}$ by assuming that $Z_1 + \cdots + Z_N = \log B$ exactly whenever $Z_1 + \cdots + Z_N \leq \log B$, and that $Z_1 + \cdots + Z_N = \log A$ exactly whenever $Z_1 + \cdots + Z_N \geq \log A$. This approximation works fairly well if $f_1(x)$ and $f_2(x)$ do not differ greatly from each other, for then the quantity $\log [f_2(X_N)/f_1(X_N)]$ will probably be close to zero and will not move the sum $Z_1 + \cdots + Z_N$ very far below $\log B$ or above $\log A$. Using this approximation, we find that if $F_1(x)$ is the true common cdf, $E\{Z_1 + \cdots + Z_N\} = [1 - \beta(A,B)] \log B + \beta(A,B) \log A$, and if $F_2(x)$ is the true common cdf, $E\{Z_1 + \cdots + Z_N\} = \alpha(A,B) \log B + [1 - \alpha(A,B)] \log A$. Now we denote $E\{Z_1\}$ when $F_i(x)$ is the true cdf by h_i, for $i = 1, 2$. Then, using (7.22) and the approximation just developed, we find approximately

$$n_1(A,B) = \frac{[1 - \beta(A,B)] \log B + \beta(A,B) \log A}{h_1}$$
$$n_2(A,B) = \frac{\alpha(A,B) \log B + [1 - \alpha(A,B)] \log A}{h_2} \qquad (7.23)$$

(It can be shown that neither h_1 nor h_2 is equal to zero.)

Finally, applying the approximations (7.19) and (7.23) to (7.16) we get the approximations

$$\begin{aligned} r(1;t_b) &= \left(\frac{A-1}{A-B}\right)a(G_1;d_1) + \left(\frac{1-B}{A-B}\right)a(G_1;d_2) \\ &\quad + \frac{c}{h_1}\left[\left(\frac{A-1}{A-B}\right)\log B + \left(\frac{1-B}{A-B}\right)\log A\right] \\ r(2;t_b) &= B\left(\frac{A-1}{A-B}\right)a(G_2;d_1) + \left(\frac{A-AB}{A-B}\right)a(G_2;d_2) \\ &\quad + \frac{c}{h_2}\left[B\left(\frac{A-1}{A-B}\right)\log B + \left(\frac{A-AB}{A-B}\right)\log A\right] \end{aligned} \qquad (7.24)$$

To find the values of A and B that give an approximately minimax decision rule, we equate the two right-hand expressions in (7.24) and use this equation to express A in terms of B, and then we choose B to minimize the right-hand sides of (7.24). The actual computation of A and B from these conditions can be quite laborious. Sometimes the use of graphs is convenient. In the next section we discuss a numerical example.

7.5. A Numerical Example of an Approximately Minimax Wald Sequential Rule. Suppose an agricultural area has been invaded by a swarm of insects, which are busily laying eggs. Unless the area is sprayed in time, the hatched larvae will devour an expected $250,000 worth of crops. However, the insects are all of variety V_1 or variety V_2, which one is not known, and only one specific spray is effective against V_1, while a different spray is effective against V_2. Also, only one spray can be used, since in combination they poison all animal life. It is impossible to tell for certain which variety of insect the invaders are until the eggs hatch (and then it is too late to spray), but the geneticists inform us that the number of spots on an internal organ of each insect is a chance variable X whose distribution depends on the variety of insect. For an insect of variety V_1, we have the distribution $f_1(x)$ given as follows:

x	0	1	2
$f_1(x)$	$1/16$	$6/16$	$9/16$

For an insect of variety V_2, we have the distribution $f_2(x)$ given as follows:

x	0	1	2
$f_2(x)$	$9/16$	$6/16$	$1/16$

It costs $1 to have the number of spots counted for each insect. The cost of spraying is negligible. The problem is to decide which type of spray to use, on the basis of observations of the numbers of spots on as many insects as we care to inspect.

We label as d_1 the decision to use the spray effective against V_1 and as d_2 the decision to use the spray effective against V_2. $X_1, X_2, \ldots$ are the numbers of spots on the first, second, . . . insect we inspect. The chance variable Y is the value of the crops that will be destroyed by the larvae. We assume that if the correct spray is chosen, it is 100 per cent effective. Then we have $a(G_1;d_1) = 0$, $a(G_1;d_2) = 250{,}000$, $a(G_2;d_1) = 250{,}000$, $a(G_2;d_2) = 0$. Also, $c = 1$, and $h_1 = (1/16) \log 9 + (6/16) \log 1 + (9/16) \log 1/9 = -1.1$, $h_2 = (9/16) \log 9 + (6/16) \log 1 + (1/16) \log 1/9 = 1.1$. Then the right-hand sides of Eq. (7.24) of Sec. 7.4 become

$$250{,}000\left(\frac{1-B}{A-B}\right) - \frac{1}{1.1}\left[\left(\frac{A-1}{A-B}\right)\log B + \left(\frac{1-B}{A-B}\right)\log A\right] \qquad (7.25)$$

$$250{,}000B\left(\frac{A-1}{A-B}\right) + \frac{1}{1.1}\left[B\left(\frac{A-1}{A-B}\right)\log B + \left(\frac{A-AB}{A-B}\right)\log A\right]$$

Equating these expressions gives $A = 1/B$. Substituting $A = 1/B$ in the first expression of (7.25) gives

$$\frac{250{,}000B}{1+B} - \left(\frac{1-B}{1+B}\right)\frac{\log B}{1.1}$$

The value of B that minimizes this last expression is approximately 1/274,975. Therefore the value of A is approximately 274,975.

Denote by $T_i(m)$ the number of the chance variables $X_1, \ldots, X_m$ which are equal to i (i can be 0, 1, or 2). Then our approximately minimax Wald sequential rule can be described as follows: Continue sampling as long as $1/274{,}975 < 9^{T_0(m) - T_2(m)} < 274{,}975$, and as soon as one of these inequalities fails to hold, stop sampling and choose d_1 if $9^{T_0(m) - T_2(m)} \leq 1/274{,}975$; choose d_2 if $9^{T_0(m) - T_2(m)} \geq 274{,}975$.

It is easily seen that in any problem where $a(G_1;d_1) = a(G_2;d_2)$, $a(G_2;d_1) = a(G_1;d_2)$, and $h_1 = -h_2$, then the A and B that give the approximately minimax Wald sequential rule are related by the equation $AB = 1$. Our example illustrates this.

7.6. Problems in Which a Sequence of Regular Decisions Must Be Made over Time. Denote by $Y_1(j), \ldots, Y_{n_j}(j)$ the chance variables that will be observed between the jth time at which we must choose a decision and the $(j + 1)$st time at which we must choose a decision. $X_1, \ldots, X_m$ denote, as usual, the chance variables that are observed before any decision must be chosen. To save space, we denote the set $Y_1(j), \ldots, Y_{n_j}(j)$ by $Y(j)$ and the set $X_1, \ldots, X_m$ by X. $D(j)$ denotes the decision made at the jth time. Suppose that a decision must be chosen at T different times. The loss depends on X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots$, $D(T)$, $Y(T)$ and will be written $W(X, D(1), Y(1), D(2), Y(2), \ldots, D(T), Y(T))$.

The key fact about the construction of a Bayes decision rule relative to a given $B(\theta)$ in this case is that we must *first* describe how the decision rule chooses $D(T)$; then we describe how the decision rule chooses $D(T - 1)$; then how the decision rule chooses $D(T - 2)$; etc. In other words, we must work our way backward in the construction of a Bayes decision rule. This is because in order to evaluate the goodness of a decision to be made at a certain time, we have to know how we are going to proceed in the future (that is, how we are going to make decisions in the future).

In choosing $D(T)$, our decision rule will of course take into account the values of X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots$, $D(T - 1)$, $Y(T - 1)$, which will be known at the time $D(T)$ will have to be chosen. Thus, for the problem of choosing $D(T)$, the quantities X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots$, $D(T - 1)$, $Y(T - 1)$ play the role that $X_1, \ldots, X_m$ played in the problems of Chap. 5, where T was equal to 1, while $Y(T)$ plays the role that $Y_1, \ldots, Y_n$ did in the problems of Chap. 5. Thus the problem of describing how a Bayes decision rule relative to $B(\theta)$ chooses $D(T)$ is a problem of the type described in Chap. 5.

After we have described how the decision rule chooses $D(T)$, we have expressed $D(T)$ in terms of X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots$, $D(T - 1)$,

$Y(T-1)$. But then for the problem of choosing $D(T-1)$, we have eliminated the quantity $D(T)$ by expressing it in terms of the quantities X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots, D(T-1)$, $Y(T-1)$. In choosing $D(T-1)$, our decision rule will take into account the values of X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots, D(T-2)$, $Y(T-2)$, which will be known at the time $D(T-1)$ will have to be chosen. Thus, for the problem of choosing $D(T-1)$, the quantities X, $D(1)$, $Y(1)$, $D(2)$, $Y(2), \ldots, D(T-2)$, $Y(T-2)$ play the role that $X_1, \ldots, X_m$ played in Chap. 5, while the quantities $Y(T-1)$, $Y(T)$ play the role that $Y_1, \ldots, Y_n$ played in Chap. 5. Thus, once we have described how the decision rule chooses $D(T)$, deciding how the decision rule chooses $D(T-1)$ is a problem of the type described in Chap. 5.

The reasoning used above can be applied to the problem of choosing $D(T-2)$, $D(T-3)$, etc. For the problem of choosing $D(j)$, we have already described how the decision rule chooses $D(T)$, $D(T-1), \ldots, D(j+1)$. But this means that we have expressed $D(T)$, $D(T-1), \ldots, D(j+1)$ in terms of X, $D(1)$, $Y(1), \ldots, D(j-1)$, $Y(j-1)$, $D(j)$, $Y(j)$, $Y(j+1), \ldots, Y(T-1)$, eliminating the quantities $D(T)$, $D(T-1), \ldots, D(j+1)$. Then the problem of choosing $D(j)$ is a problem of the type discussed in Chap. 5. The quantities X, $D(1)$, $Y(1), \ldots, D(j-1)$, $Y(j-1)$ play the role that $X_1, \ldots, X_m$ played in Chap. 5, and the quantities $Y(j)$, $Y(j+1), \ldots, Y(T)$ play the role that $Y_1, \ldots, Y_n$ played in Chap. 5.

The construction of an over-all Bayes decision rule relative to $B(\theta)$ requires T applications of the technique of Chap. 5: one for describing how $D(T)$ is to be chosen; then one for describing how $D(T-1)$ is to be chosen; . . . ; then one for describing how $D(1)$ is to be chosen.

A numerical example will illustrate the description above. Suppose a company has promised to deliver two items of a certain type to a customer by a certain date. The production of these items is a two-stage process, and in each stage there is a constant probability $1-\theta$ that the item will be spoiled during the stage. Only the half-finished items that survive the first stage can enter the second stage. Spoiled and surplus items have no value. Because of time limitations, no reruns are possible. The maximum capacity of production in each stage is 5. Before starting the first stage of the production process, the company will be able to observe the number of items surviving a single-stage production process in which the probability of spoilage of each item is $1-\theta$, out of two items started through the single-stage process. Costs are as follows. The company is to be paid \$2,000, but will pay a penalty cost of \$1,200 if it delivers only one item and a penalty cost of \$2,400 if it delivers no items. It costs the company \$300 for each item started through the first stage of production and \$100 for each item started through the second stage of production. Suppose the value of θ is unknown (except for the obvious fact that it is between 0 and 1) and we want to construct a Bayes decision rule for the problem relative to the a priori distribution $B(\theta)$ which has pdf $b(\theta) = 2(1 - \theta)$ for $0 < \theta < 1$.

We set up the following notation. X denotes the number of items surviving out of the two items started through the single-stage process that will be observed before any decisions must be chosen. $D(1)$ is the number of items started through the first stage of the two-stage process. Because of the restriction on capacity, the possible values of $D(1)$ are $0, 1, 2, 3, 4, 5$. $Y(1)$ is the number of items surviving the first stage. $D(2)$ is the number of items started through the second stage. $Y(2)$ is the number of items surviving the second stage. Clearly, $D(1) \geq Y(1) \geq D(2) \geq Y(2)$. From the description above, it is easily seen that the loss function does not depend on X, and is given as follows:

$$W(D(1), Y(1), D(2), Y(2)) = -2{,}000 + 300\,D(1) + 100\,D(2) \qquad \text{if } Y(2) \geq 2$$

$$W(D(1), Y(1), D(2), Y(2)) = -2{,}000 + 300\,D(1) + 100\,D(2) + 1{,}200[2 - Y(2)] \qquad \text{if } Y(2) < 2$$

Also, it is clear that for given values of $D(1)$ and $D(2)$, the chance variables X, $Y(1)$, and $Y(2)$ are independent, and each has a binomial distribution: the parameters for X are $2, \theta$; the parameters for $Y(1)$ are $D(1), \theta$; the parameters for $Y(2)$ are $D(2), \theta$. Thus for nonnegative integers x, $y(1)$, $y(2)$, with $x \leq 2$, $y(1) \leq D(1)$, and $y(2) \leq D(2)$, $f(x, y(1), y(2); \theta)$ is equal to the product of the following three expressions:

$$\frac{2!}{x!\,(2-x)!}\,\theta^{x}(1-\theta)^{2-x}$$

$$\frac{D(1)!}{y(1)!\,(D(1)-y(1))!}\,\theta^{y(1)}(1-\theta)^{D(1)-y(1)}$$

$$\frac{D(2)!}{y(2)!\,(D(2)-y(2))!}\,\theta^{y(2)}(1-\theta)^{D(2)-y(2)}$$

The first step in the construction of a Bayes decision rule relative to $B(\theta)$ is the description of how the rule chooses $D(2)$ for given values of X, $D(1)$, and $Y(1)$. We compute $K(D(2); x, D(1), y(1))$ exactly as in Chap. 5 and choose the value of $D(2)$ that minimizes $K(D(2); x, D(1), y(1))$ for the given values x, $D(1)$, $y(1)$. We have

$$K(D(2); x, D(1), y(1)) = \int_0^1 b(\theta)\left[\sum_{y(2)=0}^{D(2)} W(D(1), y(1), D(2), y(2))\,f(x, y(1), y(2); \theta)\right] d\theta$$

Carrying out the summation with respect to $y(2)$ gives

$$K(D(2);x,D(1),y(1)) = \int_0^1 b(\theta)\frac{2!}{x!\,(2-x)!}\,\frac{D(1)!}{y(1)!\,(D(1)-y(1))!}$$

$$\times\ \theta^{x+y(1)}(1-\theta)^{2+D(1)-x-y(1)}[-2{,}000+300\,D(1)$$

$$+\ 100\,D(2)+2{,}400(1-\theta)^{D(2)}$$

$$-\ 1{,}200\,D(2)\theta(1-\theta)^{D(2)-1}]\,d\theta$$

Setting $b(\theta) = 2(1-\theta)$ and integrating, making use of the fact that

$$\int_0^1 \theta^r(1-\theta)^s\,d\theta = \frac{r!\,s!}{(r+s+1)!}$$

for any nonnegative integers r and s, we find that $K(D(2);x,D(1),y(1))$ is given by the expression

$$2\,\frac{2!}{x!\,(2-x)!}\,\frac{D(1)!}{y(1)!\,(D(1)-y(1))!}\Big\{[-2{,}000+300\,D(1)+100\,D(2)]$$

$$\times\ \frac{(x+y(1))!\,(3+D(1)-x-y(1))!}{(4+D(1))!}$$

$$+\ 2{,}400\,\frac{(x+y(1))!\,(3+D(1)+D(2)-x-y(1))!}{(4+D(1)+D(2))!}$$

$$-\ 1{,}200\,D(2)\,\frac{(x+y(1)+1)!\,(2+D(1)+D(2)-x-y(1))!}{(4+D(1)+D(2))!}\Big\}$$

A detailed calculation and comparison shows that if $D(1) \leq 4$, then $D(2)$ should be set equal to $Y(1)$ no matter what the value of X is. If $D(1) = 5$, then if $X = 0$, we should set $D(2)$ equal to $Y(1)$, while if X is 1 or 2, then $D(2)$ should be set equal to $Y(1)$ if $Y(1) \leq 4$ and should be set equal to 4 if $Y(1)$ is equal to 5. This completely describes how the Bayes decision rule chooses $D(2)$.

Our next task is to describe how the decision rule chooses $D(1)$. To do this we compute

$$K(D(1);x) = \int_0^1 b(\theta)\left[\sum_{y(1)=0}^{D(1)}\ \sum_{y(2)=0}^{D(2)} W(D(1),y(1),D(2),y(2))f(x,y(1),y(2);\theta)\right]d\theta$$

where in the computation the value of $D(2)$ is determined by the values of x, $D(1)$, and $y(1)$, as described in the preceding paragraph. The computation of $K(D(1);x)$ is routine but lengthy. The numerical results are

$$\begin{array}{lll}
K(0;0) = 200 & K(0;1) = 133.3 & K(0;2) = 66.7 \\
K(1;0) = 320 & K(1;1) = 166.7 & K(1;2) = 46.7 \\
K(2;0) = 440 & K(2;1) = 200 & K(2;2) = 26.7 \\
K(3;0) = 563.1 & K(3;1) = 246.7 & K(3;2) = 33.3 \\
K(4;0) = 689 & K(4;1) = 305.4 & K(4;2) = 57 \\
K(5;0) = 818 & K(5;1) = 374 & K(5;2) = 89
\end{array}$$

From these values, we see that the Bayes decision rule sets $D(1) = 0$ if $X = 0$ or 1 and sets $D(1) = 2$ if $X = 2$.

In summary, a Bayes decision rule relative to $B(\theta)$ is to produce nothing if $X = 0$ or 1 and to start two items through the first stage if $X = 2$. All survivors (if any) of the first stage are started through the second stage when $D(1) = 2$.

Chapter 8

THE EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION

8.1. Introduction. In many important problems, $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ are all independently distributed, each with the same distribution, which is unknown, and the loss does not depend on the values of $X_1, \ldots, X_m$. Then we can write the loss function as $W(Y_1, \ldots, Y_n; D)$. Denote the common unknown cdf of $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ by $F(x)$. The expected value of the loss, for any fixed decision D, depends only on D and on $F(x)$, and we denote this expected loss by $S(F;D)$. If we knew $F(x)$, we would simply choose the value of D to minimize $S(F;D)$, and it would not be necessary to observe $X_1, \ldots, X_m$. It is because we do not know $F(x)$ that we observe $X_1, \ldots, X_m$ and try to estimate $F(x)$ from the observed values of $X_1, \ldots, X_m$. In this chapter we describe a certain way of estimating $F(x)$.

We define the "empirical cdf based on $X_1, \ldots, X_m$," denoted by $H(x; X_1, \ldots, X_m)$, as follows: for any given x,

$$H(x; X_1, \ldots, X_m) = \frac{\text{number of variables } X_1, \ldots, X_m \text{ no greater than } x}{m}$$

Note that this function $H(x; X_1, \ldots, X_m)$ is defined for all values of x, and has the essential properties of a cdf: it is nondecreasing and approaches 0 as x decreases and approaches 1 as x increases. We shall show that as m increases, the probability that $H(x; X_1, \ldots, X_m)$ is close to $F(x)$ for all values of x approaches 1.

8.2. Stochastic Convergence and Tchebycheff's Inequality. If $Z_1, Z_2, \ldots$ are chance variables and k is a constant, we say "Z_i converges stochastically to k as i increases" if the following is true: For each pair of positive numbers ϵ, δ, we can find a positive integer $n(\epsilon,\delta)$ such that for each and every integer m above $n(\epsilon,\delta)$, the inequality $P(|Z_m - k| \leq \epsilon) > 1 - \delta$ holds.

Speaking roughly, to say that Z_i converges stochastically to k as i increases means that for large m the probability that Z_m is close to k is almost 1.

A useful device for proving stochastic convergence is "Tchebycheff's inequality," which states that if W is a chance variable such that $P(W < 0) = 0$, then for any positive number b, $P(W < b) \geq 1 - E\{W\}/b$.

First we prove Tchebycheff's inequality for the case where W has a pdf $g(w)$. Since $P(W < 0) = 0$, $g(w) = 0$ for $w < 0$. Then

$$E\{W\} = \int_0^\infty wg(w)\,dw = \int_0^b wg(w)\,dw + \int_b^\infty wg(w)\,dw$$
$$\geq \int_b^\infty wg(w)\,dw \geq \int_b^\infty bg(w)\,dw = b\int_b^\infty g(w)\,dw = bP(W \geq b)$$

which gives $P(W \geq b) \leq E\{W\}/b$, or $P(W < b) \geq 1 - E\{W\}/b$, completing the proof.

Next we prove Tchebycheff's inequality for the case where W has a distribution which can be given in table form, as follows:

Possible values	w_1	w_2	w_3	$\cdots$
Probability	p_1	p_2	p_3	$\cdots$

where we may assume that $0 \leq w_1 < w_2 < \cdots$. Denote by k the largest integer such that $w_k < b$. Then

$$E\{W\} = \sum_{i=1}^{\infty} p_i w_i = \sum_{i=1}^{k} p_i w_i + \sum_{i=k+1}^{\infty} p_i w_i \geq \sum_{i=k+1}^{\infty} p_i w_i \geq b \sum_{i=k+1}^{\infty} p_i = bP(W \geq b)$$

which gives $P(W \geq b) \leq E\{W\}/b$, or $P(W < b) \geq 1 - E\{W\}/b$, completing the proof.

A useful application of Tchebycheff's inequality will now be described. Suppose Z_m has a binomial distribution with parameters m, p for each positive integer m. Then we show that Z_i/i converges stochastically to p as i increases. To do this, we define W_i as $((Z_i/i) - p)^2$. Then of course $P(W_i < 0) = 0$, and it is easily found that $E\{W_i\} = p(1 - p)/i$. Tchebycheff's inequality then gives $P(W_i < b) \geq 1 - (p(1 - p)/bi)$, for any positive b. But $P(W_i < b) = P(|(Z_i/i) - p| < \sqrt{b})$. If we set $\epsilon = \sqrt{b}$, then for every i greater than $p(1 - p)/\delta\epsilon^2$ we have $P(|(Z_i/i) - p| \leq \epsilon) > 1 - \delta$. This completes the proof that Z_i/i converges stochastically to p as i increases.

8.3. An Inequality on the Probability of a Combined Event. To simplify the notation, we denote the event (not A) by $\bar{A}$. Then we have the following useful inequality. If $A_1, A_2, \ldots, A_k$ are any events, the inequality $P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k) \geq 1 - P(\bar{A}_1) - P(\bar{A}_2) - \cdots - P(\bar{A}_k)$ holds.

To prove this, first we note that $P(\bar{A}_1 \text{ or } \bar{A}_2 \text{ or } \cdots \text{ or } \bar{A}_k) \leq P(\bar{A}_1) + P(\bar{A}_2) + \cdots + P(\bar{A}_k)$. This is easily seen, since the right-hand side would overcount trials on which more than one of the events $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_k$ occurs. Next we note that the event $(\bar{A}_1 \text{ or } \bar{A}_2 \text{ or } \cdots \text{ or } \bar{A}_k)$ is the same event as $\overline{(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k)}$. Therefore $P(\bar{A}_1 \text{ or } \bar{A}_2 \text{ or } \cdots \text{ or } \bar{A}_k) = P\overline{(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k)} = 1 - P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k) \leq P(\bar{A}_1) + P(\bar{A}_2) + \cdots + P(\bar{A}_k)$, and the proof follows very simply from this.

Note that if $A_1, A_2, \ldots, A_k$ all have high probabilities, then the inequality shows that the event $(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k)$ also has a high probability.

8.4. The Convergence of the Empirical cdf to the True cdf. Throughout this chapter, we are assuming that $X_1, \ldots, X_m$ are independent chance variables, each with the same cdf $F(x)$. We have defined the empirical cdf $H(x; X_1, \ldots, X_m)$ in Sec. 8.1. In this section, we shall show that for m large, $H(x; X_1, \ldots, X_m)$ will be close to $F(x)$ with high probability. We show this by means of three theorems.

Theorem 1. *For any given value b, $H(b; X_1, \ldots, X_m)$ converges stochastically to $F(b)$ as m increases.*

Proof. First we note that $mH(b; X_1, \ldots, X_m)$ is a chance variable with a binomial distribution with parameters m, $F(b)$. For $mH(b; X_1, \ldots, X_m)$ is the number of the values $X_1, \ldots, X_m$ which are no greater than b. But $X_1, \ldots, X_m$ are independent, and each has probability $F(b)$ of being no greater than b. This tells us that $mH(b; X_1, \ldots, X_m)$ has a binomial distribution with parameters m, $F(b)$. But then Theorem 1 follows immediately from the example given at the end of Sec. 8.2.

Theorem 2. *For any given positive ϵ, there is a finite number (k, say) of values $b_1 < b_2 < \cdots < b_k$, such that if $|H(b_i; X_1, \ldots, X_m) - F(b_i)| < \epsilon/2$ for $i = 1, \ldots, k$, then* $\max\limits_x |H(x; X_1, \ldots, X_m) - F(x)| < \epsilon$.

Proof. First we give the proof for the case where $F(x)$ is continuous. Then we can define b_i as a value satisfying the equation $F(b_i) = i\epsilon/2$ and k as the largest integer such that $k\epsilon/2 < 1$. Such values b_i always exist when $F(x)$ is continuous. We also define b_0 as $-\infty$ and b_{k+1} as $+\infty$, for convenience in writing, noting that

$$|H(b_0; X_1, \ldots, X_m) - F(b_0)| = 0$$

and

$$|H(b_{k+1}; X_1, \ldots, X_m) - F(b_{k+1})| = 0$$

Next we show that if

$$|H(b_i; X_1, \ldots, X_m) - F(b_i)| < \frac{\epsilon}{2}$$

and

$$|H(b_{i+1}; X_1, \ldots, X_m) - F(b_{i+1})| < \frac{\epsilon}{2}$$

then

$$|H(x; X_1, \ldots, X_m) - F(x)| < \epsilon \qquad \text{for all } x \text{ between } b_i \text{ and } b_{i+1}$$

Suppose not. Then we have either

Case 1:

$$H(x'; X_1, \ldots, X_m) \geqslant F(x') + \epsilon \qquad \text{for some } x' \text{ between } b_i \text{ and } b_{i+1}$$

Case 2:

$$H(x'; X_1, \ldots, X_m) \leqslant F(x') - \epsilon \qquad \text{for some } x' \text{ between } b_i \text{ and } b_{i+1}$$

In case 1,

$$\begin{aligned} H(x'; X_1, \ldots, X_m) &\geqslant F(x') + \epsilon \\ &\geqslant F(b_i) + \epsilon \\ &= \frac{i\epsilon}{2} + \epsilon \\ &= \frac{(i+1)\epsilon}{2} + \frac{\epsilon}{2} \\ &= F(b_{i+1}) + \frac{\epsilon}{2} \end{aligned}$$

which is impossible, since $H(x'; X_1, \ldots, X_m) \leqslant H(b_{i+1}; X_1, \ldots, X_m) < F(b_{i+1}) + \epsilon/2$. Thus case 1 cannot occur.

In case 2,

$$\begin{aligned} H(x'; X_1, \ldots, X_m) \leqslant F(x') - \epsilon \leqslant F(b_{i+1}) - \epsilon &= \frac{(i+1)\epsilon}{2} - \epsilon \\ &= \frac{i\epsilon}{2} - \frac{\epsilon}{2} \\ &= F(b_i) - \frac{\epsilon}{2} \end{aligned}$$

which is impossible, since $H(x'; X_1, \ldots, X_m) \geqslant H(b_i; X_1, \ldots, X_m) > F(b_i) - \epsilon/2$. Thus case 2 cannot occur.

The impossibility of cases 1 and 2 shows that if $|H(b_i; X_1, \ldots, X_m) - F(b_i)| < \epsilon/2$ and $|H(b_{i+1}; X_1, \ldots, X_m) - F(b_{i+1})| < \epsilon/2$, then $|H(x; X_1, \ldots, X_m) - F(x)| < \epsilon$ for all x between b_i and b_{i+1}. Theorem 2 follows immediately.

The proof for the case where $F(x)$ is not continuous is exactly the same as the proof just given, if we can find values b_i such that $F(b_i) = i\epsilon/2$ for all positive integers i, not greater than k, where k is the largest integer with $k\epsilon/2 < 1$. If we cannot find such values, it is because of the presence of discontinuities in $F(x)$ at certain inconvenient places. This situation is handled by including among the points $b_1, \ldots, b_k$ all points at which $F(x)$ jumps a distance of at least $\epsilon/2$. The details of the proof will not be carried out.

Theorem 3. $\max_x |H(x; X_1, \ldots, X_m) - F(x)|$ *converges stochastically to zero as m increases.*

Proof. We must show that for any given positive ϵ, δ, there is an integer $n(\epsilon,\delta)$ such that $P[\max_x |H(x; X_1, \ldots, X_m) - F(x)| < \epsilon] > 1 - \delta$ for any $m > n(\epsilon,\delta)$. Let A_i denote the event $|H(b_i; X_1, \ldots, X_m) - F(b_i)| < \epsilon/2$, for $i = 1, \ldots, k$, where $b_1 < b_2 < \cdots < b_k$ are the values described in Theorem 2. The event A_1 and A_2 and $\cdots$ and A_k then implies the event $\max_x |H(x; X_1, \ldots, X_m) - F(x)| < \epsilon$, so that

$$P[\max_x |H(x; X_1, \ldots, X_m) - F(x)| < \epsilon] \geq P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k).$$

By Theorem 1, there is a positive integer M such that $P(A_i) > 1 - \delta/k$ for any $m > M$, and $i = 1, \ldots, k$. Then $P(\bar{A}_i) < \delta/k$ for any $m > M$, and $i = 1, \ldots, k$. By the inequality of Sec. 8.3, $P(A_1$ and A_2 and $\cdots$ and $A_k) \geq 1 - P(\bar{A}_1) - P(\bar{A}_2) - \cdots - P(\bar{A}_k) > 1 - k(\delta/k) = 1 - \delta$, if $m > M$. Thus $P[\max_x |H(x; X_1, \ldots, X_m) - F(x)| < \epsilon] > 1 - \delta$ for any $m > M$, and this completes the proof of Theorem 3, with $n(\epsilon,\delta) = M$.

8.5. The Empirical Decision Rule. In this chapter we are discussing problems in which $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are all independent, and each has cdf $F(x)$, and the loss does not depend on $X_1, \ldots, X_m$. For any fixed decision D, the expected value of the loss depends only on D and on $F(x)$, and we have denoted this expected value by $S(F;D)$.

In many important problems, $S(F;D)$ changes only slightly if $F(x)$ changes slightly. To be more precise, if ϵ is any given positive value, there is a positive value δ such that $|S(F;D) - S(G;D)| < \epsilon$ for every cdf $G(x)$ such that $\max_x |F(x) - G(x)| < \delta$. [Here δ depends on ϵ, but not on $F(x)$ or D.] In such a problem, the following decision rule seems reasonable: Choose the value of D that minimizes $S[H(x; X_1, \ldots, X_m); D]$. This decision rule will be called the "empirical decision rule." It seems reasonable because in Sec. 8.4 we showed that with a high probability $\max_x |H(x; X_1, \ldots, X_m) - F(x)|$ will be small, and therefore by the assumption just made, $S[H(x; X_1, \ldots, X_m); D]$ will be close to $S(F; D)$

for all D. The best decision is the one that minimizes $S(F;D)$, but since $F(x)$ is not known, we use $H(x; X_1, \ldots, X_m)$ in its place as an approximation.

We conclude this section with an example representative of an important class of problems. A newspaper vendor must decide how many copies of a monthly magazine to buy from the publisher. He may buy any number from 0 to C, where C is a given positive integer. He pays the publisher w dollars per copy, sells to customers at r dollars per copy, and returns unsold copies to the publisher, receiving s dollars per unsold copy returned. We assume $s < w < r$. Y denotes the number of magazines that will be requested during the coming month, and $X_1, X_2, \ldots, X_m$ denote the numbers that were requested during the m preceding months. We assume that $X_1, \ldots, X_m$, Y are independent chance variables with a common but unknown cdf $F(x)$. We denote $P(X = x)$ by $f(x)$, where of course $f(x) = 0$ unless x is a nonnegative integer. D denotes the number of copies the vendor will order from the publisher. Then the loss function is given as follows:

$$\begin{aligned} W(Y;D) &= wD - rY - s(D - Y) && \text{if } Y \leqslant D \\ W(Y;D) &= wD - rD && \text{if } Y > D \end{aligned}$$

This simplifies to

$$\begin{aligned} W(Y;D) &= (w - s)D + (s - r)Y && \text{if } Y \leqslant D \\ W(Y;D) &= (w - r)D && \text{if } Y > D \end{aligned}$$

Then we find

$$\begin{aligned} S(F;D) &= \sum_{y=0}^{D} [(w - s)D + (s - r)y]f(y) + \sum_{y=D+1}^{\infty} (w - r)Df(y) \\ &= (w - s)DF(D) + (s - r)\sum_{y=0}^{D} yf(y) + (w - r)D[1 - F(D)] \\ &= (w - r)D + (r - s)DF(D) + (s - r)\sum_{y=0}^{D} y[F(y) - F(y - 1)] \end{aligned}$$

From this last expression, it is easily seen that $S(F;D)$ changes only slightly if $F(x)$ changes slightly. Therefore it seems reasonable to use the empirical decision rule for this problem.

In order to find which decision is chosen by the empirical decision rule, we investigate the shape of the function $S(F;D)$. Computing the difference $S(F; D + 1) - S(F;D)$, we get

$$\begin{aligned} S(F; D + 1) - S(F;D) &= w - r + (r - s)[(D + 1)F(D + 1) - DF(D)] \\ &\qquad + (s - r)\{(D + 1)[F(D + 1) - F(D)]]\} \\ &= w - r - (s - r)F(D) \end{aligned}$$

From the last expression, it is easily seen that $S(F; D + 1) - S(F;D)$ is negative if $F(D) < (r - w)/(r - s)$, is zero if $F(D) = (r - w)/(r - s)$, and is positive if $F(D) > (r - w)/(r - s)$. This means that the function $S(F;D)$ is minimized by setting $D = Q + 1$, where Q is the largest positive integer such that $F(Q) < (r - w)/(r - s)$. Since the empirical decision rule acts as though the unknown $F(x)$ were exactly equal to the known empirical cdf $H(x; X_1, \ldots, X_m)$, and since the largest possible decision is equal to C, the empirical decision rule chooses a decision as follows: set $D = \min(C, Q' + 1)$, where Q' is the largest positive integer such that $H(Q'; X_1, \ldots, X_m) < (r - w)/(r - s)$.

8.6. The Empirical Decision Rule and Bayes Decision Rules. In our discussion in this chapter, we have not yet mentioned admissible decision rules or Bayes decision rules, a fact which should seem puzzling. In this section, we attempt to explain this fact.

In the whole discussion preceding this chapter, the unknown joint distribution was either one of a given finite number of possible distributions, or else one of a given family of distributions generated by the variation of a finite number of parameters. In such cases, we defined Bayes decision rules and knew that each admissible decision rule is either a Bayes decision rule or else a limit of Bayes decision rules. However, in the present chapter, we did not assume that the unknown distribution is one of a given finite number of possible distributions, or one of a given family of distributions generated by the variation of a finite number of parameters. With such a wide variety of possible distributions as we are allowing in the present chapter, it is not known whether each admissible decision rule is a Bayes decision rule or a limit of Bayes decision rules. Since the standard method for finding admissible decision rules may not work in this case, we employ a decision rule which has a certain intuitive appeal.

Chapter 9

CONVENTIONAL STATISTICAL THEORY

9.1. Introduction. In Sec. 5.3 we pointed out that the standard formulation of statistical problems has the loss depending on θ, D, and $X_1, \ldots, X_m$ and that $Y_1, \ldots, Y_n$ are not mentioned in the problem. All the techniques that we have developed for finding Bayes and admissible decision rules apply to the standard formulation, once some simple changes in notation have been made. We outline these changes in notation.

The loss function in the standard problem is written as $W(\theta;D;x)$, since $Y_1, \ldots, Y_n$ do not appear in the problem.

When the joint distribution of $X_1, \ldots, X_m$ corresponding to θ allows us to list the possible values of the chance variables, $f(x;\theta)$ denotes the probability assigned to the m-dimensional point x by this distribution. When the distribution corresponding to θ has a joint pdf, $f(x;\theta)$ denotes the value of this joint pdf at the m-dimensional point x.

If there is a finite number L of possible decisions in our problem and the possible values of $X_1, \ldots, X_m$ can be listed, then

$$r(\theta;s) = \sum_x \sum_{D=1}^{L} W(\theta;D;x)f(x;\theta)s(D;x)$$

If $X_1, \ldots, X_m$ have a joint pdf, then

$$r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{D=1}^{L} W(\theta;D;x)f(x;\theta)s(D;x)\, dx_1 \cdots dx_m$$

The quantity $K(D;x)$ introduced in Sec. 5.9 becomes in the present case $\sum_{\theta=1}^{h} b(\theta)W(\theta;D;x)f(x;\theta)$, and is used in the construction of a Bayes decision rule relative to $b(1), \ldots, b(h)$ in exactly the same way as in Sec. 5.9.

The quantity $K(D;x)$ introduced in Sec. 5.10 becomes

$$\int b(\theta)W(\theta;D;x)f(x;\theta)\,d\theta$$

and is used in exactly the same way as in Sec. 5.10.

In concluding this section, we describe a method for turning a decision problem of the type discussed in Chap. 5 (where the loss depended on $Y_1, \ldots, Y_n$) into a problem of the type discussed in this chapter (where the loss depends on θ and not on $Y_1, \ldots, Y_n$). We carry out the details only for the case where the possible values of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ can be listed, and there is a finite number L of possible decisions. Other cases can be handled similarly. From Sec. 5.6,

$$r(\theta;s) = \sum_x \sum_y \sum_{D=1}^{L} W(y;D;x)f(x,y;\theta)s(D;x)$$

We denote $\sum_y f(x,y;\theta)$ by $f(x;\theta)$ and note that $f(x;\theta)$ is the marginal distribution for $X_1, \ldots, X_m$ given by the distribution $f(x,y;\theta)$. Also, denote

$$\frac{\sum_y W(y;D;x)f(x,y;\theta)}{f(x;\theta)}$$

by $\overline{W}(\theta;D;x)$. Then we have

$$r(\theta;s) = \sum_x \sum_{D=1}^{L} \overline{W}(\theta;D;x)f(x;\theta)s(D;x)$$

But this is the expected loss for a problem in which the loss function is $\overline{W}(\theta;D;x)$ and the possible distributions for $X_1, \ldots, X_m$ are given by $f(x;\theta)$ as θ varies. But this is a problem of the type discussed in this chapter, where $Y_1, \ldots, Y_n$ do not appear.

9.2. Testing a Hypothesis. A very common type of conventional statistical problem, known as the "problem of testing a hypothesis," will now be described.

There are two possible decisions, which we label as 1, 2. The possible values of θ are broken into three nonoverlapping groups, which we label I, II, III. The loss does not depend on the values of $X_1, \ldots, X_m$, so we can write the loss function as $W(\theta;D)$. The value of $W(\theta;D)$ depends on θ only through the group in which θ falls, and is given by the following table:

		θ in		
		I	II	III
D	1	0	1	0
	2	1	0	0

Clearly, if θ is in group I, we should like to choose decision 1; if θ is in group II, we should like to choose decision 2; if θ is in group III, it does not matter which decision we choose. As a matter of terminology, choosing decision 1 is called "accepting the hypothesis that θ is in group I" and choosing decision 2 is called "rejecting the hypothesis that θ is in group I."

It is easily seen that $r(\theta;s)$ has the following properties: if θ is in group I, $r(\theta;s) = P$(decision 2 chosen by decision rule s); if θ is in group II, $r(\theta;s) = P$(decision 1 chosen by decision rule s); if θ is in group III, $r(\theta;s) = 0$.

We show this for the case where the possible values of $X_1, \ldots, X_m$ can be listed. Then

$$r(\theta;s) = \sum_x \sum_{D=1}^{2} W(\theta;D)f(x;\theta)s(D;x)$$

Then if θ is in group I, $W(\theta;1) = 0$, $W(\theta;2) = 1$, so we have

$$r(\theta;s) = \sum_x f(x;\theta)s(2;x)$$

and it is easily seen that this last sum represents the probability that decision 2 will be chosen when the decision rule s is used. The demonstrations for θ in group II and for θ in group III are entirely similar.

Usually there is an additional aspect to the problem of testing a hypothesis. There is a preassigned value $\alpha (0 < \alpha < 1)$, and we are limited to the use of a decision rule s with the property that $r(\theta;s) \leq \alpha$ for all θ in group I. The quantity α is known as the "level of significance." Of all decision rules satisfying this requirement, we should like to find a decision rule which makes $\max\limits_{\theta \text{ in II}} r(\theta;s)$ as small as possible. Such a decision rule is called a "minimax test of level of significance α."

In problems where there is a finite number of distributions in both I and II and a finite number of possible sets of values of $X_1, \ldots, X_m$ under each distribution, a minimax test of level of significance α can be constructed by the use of linear programming. The linear programming problem is set up as follows. The unknowns are $s(D;x)$ for $D = 1, 2$ and for all possible values of x and $\max\limits_{\theta \text{ in II}} r(\theta;s)$. The equalities and inequalities are

$$\begin{array}{ll} s(1;x) + s(2;x) = 1 & \text{for each } x \\ \sum_x f(x;\theta)s(2;x) \leq \alpha & \text{for each } \theta \text{ in I} \\ \sum_x f(x;\theta)s(1;x) - \max\limits_{\theta \text{ in II}} r(\theta;s) \leq 0 & \text{for each } \theta \text{ in II} \end{array}$$

and it is desired to find the values of the unknowns which minimize $\max_{\theta \text{ in II}} r(\theta;s)$.

As a numerical example, we take a case where $m = 1$, group I consists of only one distribution, and group II consists of two distributions. The distribution in group I is the binomial distribution with parameters 3, 0.4. The distributions in group II are the binomial distributions with parameters 3, 0.3, and 3, 0.5, respectively. The value of α is 0.10. It is not difficult to see that in any problem where there is only one distribution in group I, the inequality $r(\theta;s) \leq \alpha$ for the single θ in group I may be replaced by the exact equality $r(\theta;s) = \alpha$ for the single θ in group I. This is so because if $r(\theta;s)$ were below α for the θ in group I, we could raise $s(2;x)$ to make $r(\theta;s) = \alpha$, and since $s(1;x)$ would be lowered, the change would not increase our objective function. Then our linear programming equalities for the present problem are

$$s(1;0) + s(2;0) = 1$$

$$s(1;1) + s(2;1) = 1$$

$$s(1;2) + s(2;2) = 1$$

$$s(1;3) + s(2;3) = 1$$

$$0.216s(2;0) + 0.432s(2;1) + 0.288s(2;2) + 0.064s(2;3) = 0.10$$

$$0.343s(1;0) + 0.441s(1;1) + 0.189s(1;2) + 0.027s(1;3) + z_1 - \max_{\theta \text{ in II}} r(\theta;s) = 0$$

$$0.125s(1;0) + 0.375s(1;1) + 0.375s(1;2) + 0.125s(1;3) + z_2 - \max_{\theta \text{ in II}} r(\theta;s) = 0$$

where z_1 and z_2 are nonnegative slack variables. As a start, we set $s(1;0) = s(1;1) = 1$, $s(2;3) = 1$, $s(2;2) = 0.125$ (this last value is to make the fifth equality in our list hold), $\max_{\theta \text{ in II}} r(\theta;s) = 0.949$, $z_1 = 0$, $z_2 = 0.121$. Our first tableau is then

		$s(2;0)$	$s(2;1)$	$s(1;3)$	z_1
$s(1;0)$	1	−1	0	0	0
$s(1;1)$	1	0	−1	0	0
$s(2;3)$	1	0	0	−1	0
$s(1;2)$	0.875	0.75	1.5	−0.222	0
$s(2;2)$	0.125	0.75	−1.5	0.222	0
z_2	0.121	−0.357	−0.345	−0.057	1
$\max_{\theta \text{ in II}} r(\theta;s)$	0.949	−0.201	−0.157	−0.015	1

Raising $s(2;0)$, we get as our second tableau

		$s(2;2)$	$s(2;1)$	$s(1;3)$	z_1
$s(2;0)$	0.167	−1.333	−2	0.296	0
$s(1;0)$	0.833	1.333	2	−0.296	0
$s(1;1)$	1	0	−1	0	0
$s(2;3)$	1	0	0	−1	0
$s(1;2)$	1	−1	0	0	0
z_2	0.062	0.477	−0.369	−0.163	1
$\max\limits_{\theta \text{ in II}} r(\theta;s)$	0.916	0.268	0.245	−0.074	1

Raising $s(1;3)$, we get as our third tableau

		z_2	$s(2;2)$	$s(2;1)$	z_1
$s(1;3)$	0.381	−6.13	2.93	−2.26	6.13
$s(2;0)$	0.279	−1.816	−0.466	−2.67	1.816
$s(1;0)$	0.721	1.816	0.466	2.67	−1.816
$s(1;1)$	1	0	0	−1	0
$s(2;3)$	0.619	6.31	−2.93	2.26	−6.13
$s(1;2)$	1	0	−1	0	0
$\max\limits_{\theta \text{ in II}} r(\theta;s)$	0.888	0.454	0.051	0.412	0.546

The fact that the last row of this third tableau contains only positive entries tells us that it represents a solution to our problem. Thus a minimax test of level of significance 0.1 for this problem is given by setting $s(1;0) = 0.721$, $s(1;1) = 1$, $s(1;2) = 1$, $s(1;3) = 0.381$.

9.3. Testing a One-sided Hypothesis. A very common special type of hypothesis testing problem is where the possible joint distributions are given by the variation of a single parameter, denoted by θ, and group I consists of all distributions given by values of θ less than or equal to A, group II consists of all distributions given by values of θ greater than or equal to B, and group III consists of all distributions given by values of θ between A and B. Here A and B are given constants with $A < B$. This problem is called the "problem of testing a one-sided hypothesis."

We shall discuss the case where a single chance variable Z is sufficient for the decision problem and the distribution of Z has the following property. Let $g(z;\theta)$ denote the pdf of Z when the parameter equals θ

(if Z has a pdf), or denote $P(Z = z)$ when the parameter equals θ if the possible values of Z can be listed. For each and every pair of values θ_1, θ_2 with $\theta_1 < \theta_2$, the ratio $g(z;\theta_2)/g(z;\theta_1)$ increases as z increases.

We are going to show that for any problem of testing a one-sided hypothesis in which the conditions of the preceding paragraph are satisfied, the following decision rule s is a minimax test of level of significance α: Choose $D = 1$ if $Z < c$, choose $D = 2$ if $Z > c$, assign probability p to choosing $D = 1$ if $Z = c$, where c and p are quantities chosen to make $r(A;s) = \alpha$. [If Z is a continuous chance variable, $P(Z = c) = 0$, so the decision rule s is simplified by the elimination of p.]

Before proving that s is minimax, we list several examples where our conditions hold:

1. $X_1, X_2, \ldots, X_m$ are independent, each with a binomial distribution with the same parameters n, θ, where n is a known positive integer, and θ is an unknown quantity between 0 and 1. Here we have

$$f(x_1, \ldots, x_m; \theta) = \frac{n!}{x_1!\,(n - x_1)!}\,\frac{n!}{x_2!\,(n - x_2)!} \cdots \frac{n!}{x_m!\,(n - x_m)!} \times \theta^{\,x_1+\cdots+x_m}(1 - \theta)^{nm-(x_1+\cdots+x_m)}$$

Then we see that $Z = X_1 + \cdots + X_m$ is sufficient for the decision problem. Z has a binomial distribution with parameters nm, θ, so

$$g(z;\theta) = \frac{(nm)!}{z!\,(nm - z)!}\,\theta^z(1 - \theta)^{nm-z}$$

Then

$$\frac{g(z;\theta_2)}{g(z;\theta_1)} = \left(\frac{\theta_2}{\theta_1}\right)^z\left(\frac{1 - \theta_2}{1 - \theta_1}\right)^{nm-z} = \left(\frac{1 - \theta_2}{1 - \theta_1}\right)^{nm}\left(\frac{\theta_2 - \theta_1\theta_2}{\theta_1 - \theta_1\theta_2}\right)^z$$

and since $\theta_2 > \theta_1$, it is easily seen that this last expression increases as z increases.

2. $X_1, X_2, \ldots, X_m$ are independent, each with a normal distribution with the same mean θ and standard deviation σ, σ being known, θ unknown. Then

$$f(x_1, \ldots, x_m; \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m e^{-1/2\sigma^2 \sum_{i=1}^{m}(x_i - \theta)^2}$$

Denoting $(x_1 + \cdots + x_m)/m$ by z, we find that $f(x_1, \ldots, x_m; \theta)$ can be written as

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m e^{-(1/2\sigma^2)\Sigma(x_i - z)^2} e^{-(m/2\sigma^2)(z-\theta)^2}$$

From this, we see that $Z = (X_1 + \cdots + X_m)/m$ is sufficient for the decision problem. Z has a normal distribution with parameters θ, $\sigma/\sqrt{m}$, so

$$g(z;\theta) = \frac{\sqrt{m}}{\sigma\sqrt{2\pi}}\, e^{-(m/2\sigma^2)(z-\theta)^2}$$

Then

$$\frac{g(z;\theta_2)}{g(z;\theta_1)} = e^{-(m/2\sigma^2)(z-\theta_2)^2+(m/2\sigma^2)(z-\theta_1)^2} = e^{(m/2\sigma^2)(\theta_1^2-\theta_2^2)}e^{(m/\sigma^2)(\theta_2-\theta_1)z}$$

and since $\theta_2 > \theta_1$, it is easily seen that this last expression increases as z increases.

3. $X_1, X_2, \ldots, X_m$ are independent, each with a Poisson distribution with the same parameter θ. Then

$$f(x_1, \ldots, x_m; \theta) = \frac{e^{-m\theta}\theta^{x_1+\cdots+x_m}}{x_1!\, x_2! \cdots x_m!}$$

and we see that $Z = X_1 + \cdots + X_m$ is sufficient for the decision problem. Z has a Poisson distribution with parameter $m\theta$, so $g(z;\theta) = e^{-m\theta}(m\theta)^z/z!$. Then $g(z;\theta_2)/g(z;\theta_1) = e^{-m\theta_2+m\theta_1}(\theta_2/\theta_1)^z$, and since $\theta_2 > \theta_1$, it is easily seen that this last expression increases as z increases.

Now we turn to the proof that the decision rule s described above is minimax. For simplicity, we carry out the details only for the case where Z has a pdf. Let $G(z;\theta)$ denote the cdf for Z when the parameter is equal to θ. First we show that $G(z;\theta)$ does not increase as θ increases, for any fixed z. Suppose the contrary. Then there would be values z, θ_1, θ_2, with $\theta_1 < \theta_2$, such that $G(z;\theta_1) < G(z;\theta_2)$. There are two possible cases:

1. $g(z;\theta_1) > g(z;\theta_2)$
2. $g(z;\theta_1) \leq g(z;\theta_2)$

In case 1, $g(w;\theta_2) < g(w;\theta_1)$ for all $w < z$, because $g(z;\theta_2)/g(z;\theta_1)$ increases as z increases. But then

$$\int_{-\infty}^{z} g(w;\theta_2)\, dw < \int_{-\infty}^{z} g(w;\theta_1)\, dw \qquad \text{or} \qquad G(z;\theta_2) < G(z;\theta_1)$$

contradicting our assumption that $G(z;\theta_1) < G(z;\theta_2)$. In case 2, $g(w;\theta_2) \geq g(w;\theta_1)$ for all $w > z$, and then

$$\int_{z}^{\infty} g(w;\theta_2)\, dw \geq \int_{z}^{\infty} g(w;\theta_1)\, dw \qquad \text{or} \qquad 1 - G(z;\theta_2) \geq 1 - G(z;\theta_1)$$

$$\text{or} \qquad G(z;\theta_1) \geq G(z;\theta_2)$$

contradicting our assumption that $G(z;\theta_2) > G(z;\theta_1)$. Since the assumption that $G(z;\theta_2) > G(z;\theta_1)$ leads to a contradiction, we have shown that $G(z;\theta)$ does not increase as θ increases, for any fixed z.

Our next step in showing that s is minimax is to construct a Bayes decision rule s_b relative to the a priori distribution $B(\theta)$ which assigns probability b to the point $\theta = A$ and probability $1 - b$ to the point $\theta = B$, thus assigning zero probability to all other values of θ. Using the loss function given in Sec. 9.2, we find $K(1;z) = (1 - b)g(z;B)$, $K(2;z) = bg(z;A)$. Then the Bayes decision rule chooses $D = 1$ if $K(1;z) < K(2;z)$ and chooses $D = 2$ if $K(1;z) > K(2;z)$. This is equivalent to saying that the Bayes decision rule s_b chooses $D = 1$ if $g(z;B)/g(z;A) < b/(1 - b)$ and chooses $D = 2$ if $g(z;B)/g(z;A) > b/(1 - b)$. Since we are assuming that $g(z;B)/g(z;A)$ increases as z increases, this is the same as saying that s_b chooses $D = 1$ if $Z < c(b)$ and chooses $D = 2$ if $Z > c(b)$, where $c(b)$ is a quantity depending only on b. Clearly, $c(b)$ increases as b increases. We have $r(A;s_b) = 1 - G(c(b);A)$, and we can find a value of b, say, b', such that $1 - G(c(b');A) = \alpha$. Then we show that $s_{b'}$ is a minimax test of level of significance α, as follows. For any θ in group I, $r(\theta;s_{b'}) = 1 - G(c(b');\theta)$. Since A is the largest value of θ in group I, we have $G(c(b');\theta) \geq G(c(b');A)$, or $1 - G(c(b');\theta) \leq 1 - G(c(b');A) = \alpha$ for any θ in group I. Thus $s_{b'}$ does have level of significance α. For any θ in group II, $r(\theta;s_{b'}) = G(c(b');\theta)$. Since B is the smallest value of θ in group II, $r(\theta;s_{b'}) \leq r(B;s_{b'})$ for any θ in group II; that is,

$$r(B;s_{b'}) = \max_{\theta \text{ in II}} r(\theta;s_{b'})$$

Now suppose that $s_{b'}$ were not a minimax test of level of significance α. Then there would be a decision rule t, with $r(A;t) \leq \alpha$ and $r(B;t) < r(B;s_{b'})$. But we must have $b'r(A;s_{b'}) + (1 - b')r(B;s_{b'}) \leq b'r(A;t) + (1 - b')r(B;t)$, and this could happen only if $b' = 1$. But if $b' = 1$, $s_{b'}$ would always choose $D = 1$, and therefore α would be equal to zero. We are assuming that α is positive, and therefore $s_{b'}$ is minimax. $s_{b'}$ is the same decision rule as the decision rule s described earlier in this section.

As a numerical example, suppose $m = 4$ and X_1, X_2, X_3, X_4 are independent, each with a normal distribution with standard deviation equal to 1 and unknown mean θ. $A = 2$, $B = 6$, and $\alpha = 0.10$. Defining Z as $(X_1 + X_2 + X_3 + X_4)/4$, we know that the minimax test of level of significance 0.1 chooses $D = 1$ when $Z < c$, where c is chosen to make $r(2;s) = 0.1$. But when $\theta = 2$, Z has a normal distribution with mean equal to 2 and standard deviation equal to $\frac{1}{2}$. Thus c satisfies the equation

$$\frac{2}{\sqrt{2\pi}} \int_c^\infty e^{-2(z-2)^2}\, dz = 0.1$$

Making the transformation $y = (z - 2)/0.5$, we get

$$\frac{1}{\sqrt{2\pi}} \int_{(c-2)/0.5}^{\infty} e^{-(\frac{1}{2})y^2} dy = 0.1$$

Table 1 in the Appendix shows that $(c - 2)/0.5$ must be equal to 1.28, so that $c = 2.64$. Note that the value of c would not be changed by a change in B.

9.4. Testing a Two-sided Hypothesis. Another common special type of hypothesis testing problem is where the possible joint distributions are given by the variation of a single parameter, denoted by θ, and group I consists of the single distribution given by $\theta = A$, group II consists of all the distributions given by values of θ less than or equal to B_1 plus all the distributions given by values of θ greater than or equal to B_2, while group III consists of all distributions given by values of θ between B_1 and B_2 excluding the value A. Here A, B_1, B_2 are given values with $B_1 < A < B_2$. This problem is called the "problem of testing a two-sided hypothesis."

In many important cases, if we construct a decision rule s which is Bayes relative to the a priori distribution $B(\theta)$ which assigns probability b_1 to the point $\theta = B_1$, probability b_2 to the point $\theta = B_2$, and probability $1 - b_1 - b_2$ to the point $\theta = A$, where b_1 and b_2 are chosen so that $r(B_1;s) = r(B_2;s)$ and $r(A;s) = \alpha$, we find that s is a minimax test. We illustrate this for the case where $X_1, X_2, \ldots, X_m$ are independent, and each has a normal distribution with known standard deviation σ and unknown mean θ. We know that $Z = (1/m)(X_1 + \cdots + X_m)$ is sufficient for this problem, and that Z has a normal distribution with standard deviation $\sigma/\sqrt{m}$ and mean θ. Using the a priori distribution $B(\theta)$ described, we find

$$K(1;z) = b_1 \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} e^{-(m/2\sigma^2)(z-B_1)^2} + b_2 \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} e^{-(m/2\sigma^2)(z-B_2)^2}$$

$$K(2;z) = (1 - b_1 - b_2) \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} e^{-(m/2\sigma^2)(z-A)^2}$$

A detailed but straightforward investigation of the ratio $K(1;z)/K(2;z)$ shows that for any given values c_1, c_2 with $c_1 < c_2$, there are values of b_1, b_2 with $b_1 > 0$, $b_2 > 0$, $1 - b_1 - b_2 > 0$ such that $K(1;z)/K(2;z) < 1$ if $c_1 < z < c_2$ and $K(1;z)/K(2;z) > 1$ if $z < c_1$ or if $z > c_2$. Then the decision rule $s(c_1,c_2)$ which chooses $D = 1$ if $c_1 < Z < c_2$ and $D = 2$ for other values of Z is Bayes relative to $B(\theta)$ for properly chosen b_1, b_2, where $b_1 > 0$, $b_2 > 0$, $1 - b_1 - b_2 > 0$. For any θ in group II,

$$r(\theta;s(c_1,c_2)) = \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} \int_{c_1}^{c_2} e^{-(m/2\sigma^2)(z-\theta)^2}\, dz$$

We now show that this last expression, as a function of θ, increases as θ

increases from $-\infty$ to $(\frac{1}{2})(c_1 + c_2)$ and decreases as θ increases from $(\frac{1}{2})(c_1 + c_2)$ to ∞. This is shown by the fact that

$$\frac{\partial}{\partial\theta}\int_{c_1}^{c_2} e^{-(m/2\sigma^2)(z-\theta)^2}\,dz = \int_{c_1}^{c_2} e^{-(m/2\sigma^2)(z-\theta)^2}\,\frac{m}{\sigma^2}(z-\theta)\,dz$$

$$= \frac{1}{\sigma}\left[e^{-(m/2)[(c_1-\theta)/\sigma]^2} - e^{-(m/2)[(c_2-\theta)/\sigma]^2}\right]$$

and this derivative is zero for $\theta = (\frac{1}{2})(c_1 + c_2)$, positive for $\theta < (\frac{1}{2})(c_1 + c_2)$, and negative for $\theta > (\frac{1}{2})(c_1 + c_2)$. Next let $c_1{}'$, $c_2{}'$ be the values such that $r(B_1;s(c_1{}',c_2{}')) = r(B_2;s(c_1{}',c_2{}'))$ and $r(A;s(c_1{}',c_2{}')) = \alpha$. Such values $c_1{}'$, $c_2{}'$ always exist. Now we can show that $s(c_1{}',c_2{}')$ is a minimax test of level of significance α. First, it is clear that

$$r(B_1;s(c_1{}',c_2{}')) = r(B_2;s(c_1{}',c_2{}')) = \max_{\theta \text{ in } \Pi} r(\theta;s(c_1{}',c_2{}'))$$

because of the shape of $r(\theta;s(c_1{}',c_2{}'))$ that we established above. Now suppose that $s(c_1{}',c_2{}')$ is not minimax. Then there would be a decision rule t with

$$r(A;t) \le \alpha, \quad r(B_1;t) < r(B_1;s(c_1{}',c_2{}')), \quad \text{and} \quad r(B_2;t) < r(B_2;s(c_1{}',c_2{}'))$$

We know that $s(c_1{}',c_2{}')$ is a Bayes decision rule relative to $B(\theta)$ for properly chosen values b_1, b_2, with $b_1 > 0$, $b_2 > 0$, $1 - b_1 - b_2 > 0$. Then we must have

$$b_1r(B_1;s(c_1{}',c_2{}')) + b_2r(B_2;s(c_1{}',c_2{}')) + (1 - b_1 - b_2)r(A;s(c_1{}',c_2{}'))$$

$$\le b_1r(B_1;t) + b_2r(B_2;t) + (1 - b_1 - b_2)r(A;t)$$

This inequality contradicts the assumption that $r(A;t) \le \alpha$, $r(B_1;t) < r(B_1;s(c_1{}',c_2{}'))$, $r(B_2;t) < r(B_2;s(c_1{}',c_2{}'))$. The contradiction proves that $s(c_1{}',c_2{}')$ is minimax.

Before taking up a numerical example, we note that $c_1{}'$, $c_2{}'$ will be symmetrically placed around B_1, B_2; that is, $c_1{}' = (\frac{1}{2})(B_1 + B_2) - d$, $c_2{}' = (\frac{1}{2})(B_1 + B_2) + d$, where d is some positive value chosen to make $r(A;s(c_1{}',c_2{}')) = \alpha$. Suppose we set $m = 16$, $\sigma = 2$, $A = 3$, $B_1 = 2$, $B_2 = 5$, $\alpha = 0.05$.

Then $c_1{}' = 3.5 - d$, $c_2{}' = 3.5 + d$, where d must be chosen so that

$$r(3;s(c_1{}',c_2{}')) = 1 - \frac{4}{2\sqrt{2\pi}}\int_{c_1{}'}^{c_2{}'} e^{-2(z-3)^2}\,dz = 0.05$$

Making the transformation $y = 2(z - 3)$, we find that d satisfies the equation

$$\frac{1}{\sqrt{2\pi}}\int_{1-2d}^{1+2d} e^{-(y^2/2)}\,dy = 0.95$$

By trial and error in Table 1 in the Appendix, we find that d is approximately 1.32. Thus $c_1' = 2.18$, $c_2' = 4.82$.

9.5. Point Estimation. Another very common type of conventional statistical problem is the problem of "point estimation," which will be described in this section.

The possible joint distributions are given by the variation of a single parameter θ over a given.interval, which may be an infinite interval. The possible decisions are the possible values of θ. Standard discussions of the problem of point estimation make it clear that we should like the decision to be close to the true value of θ, but usually do not specify a loss function. However, the following type of loss function seems to be suitable, in the light of most discussions of point estimation:

$$W(\theta; D) = c(\theta)(D - \theta)^2$$

where $c(\theta)$ is a function of θ which is never negative.

With the type of loss function introduced in the preceding paragraph, suppose we want to construct a Bayes decision rule relative to an a priori distribution $B(\theta)$ with pdf $b(\theta)$. Then

$$K(D;x) = \int b(\theta)c(\theta)(D - \theta)^2 f(x_1, \ldots, x_m; \theta)\, d\theta$$

Expanding $(D - \theta)^2$, we find

$$\begin{aligned} K(D;x) = D^2 &\int b(\theta)c(\theta)f(x_1, \ldots, x_m;\ \theta)\ d\theta \\ - 2D &\int \theta b(\theta)c(\theta)f(x_1, \ldots, x_m; \theta)\ d\theta \\ + &\int \theta^2 b(\theta)c(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta \end{aligned}$$

The Bayes decision rule chooses the value of D that minimizes $K(D;x)$, and by solving the equation $(\partial/\partial D)K(D;x) = 0$, it is easily found that the minimizing value of D is

$$\frac{\int \theta b(\theta)c(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta}{\int b(\theta)c(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta}$$

As an example of the computation described in the preceding paragraph, we take the case where $X_1, \ldots, X_m$ are independent, each with a normal distribution with the same parameters: known standard deviation σ and unknown mean θ. $c(\theta) = 1$, so that $W(\theta; D) = (D - \theta)^2$. Suppose we want to construct a Bayes decision rule s_v relative to the a priori distribution with pdf $b_v(\theta) = (1/v\sqrt{2\pi})e^{-(\theta^2/2v^2)}$. Defining Z as $(1/m)(X_1 + \cdots + X_m)$, we know that Z is sufficient for the decision problem. Z has a normal distribution with standard deviation $\sigma/\sqrt{m}$ and mean θ, so the pdf for Z is $(\sqrt{m}/\sigma\sqrt{2\pi}) \exp[-(m/2\sigma^2)(z - \theta)^2]$. Then, using the formula given in the preceding paragraph, we find that s_v chooses the decision

$$\frac{\displaystyle\int_{-\infty}^{\infty} \frac{\theta}{v\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2v^2}\right) \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{m}{2\sigma^2}(z - \theta)^2\right) d\theta}{\displaystyle\int_{-\infty}^{\infty} \frac{1}{v\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2v^2}\right) \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{m}{2\sigma^2}(z - \theta)^2\right) d\theta}$$

After canceling common constant factors in numerator and denominator, combining exponents, and making the change of variable $u = \theta\sqrt{1/v^2 + m/\sigma^2}$, we can write the decision chosen by s_v as

$$\frac{1}{\sqrt{\dfrac{1}{v^2} + \dfrac{m}{\sigma^2}}} \; \frac{\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} u e^{-u^2/2} e^{Au} \, du}{\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} e^{-u^2/2} e^{Au} \, du} \tag{9.1}$$

where A denotes

$$\frac{mz}{\sigma^2 \sqrt{\dfrac{1}{v^2} + \dfrac{m}{\sigma^2}}}$$

But the integral in the denominator of (9.1) has been evaluated in Sec. 4.6, where we found

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} e^{Au} \, du = e^{(1/2)A^2}$$

Also, the integral in the numerator of (9.1) is equal to the derivative with respect to A of the integral in the denominator of (9.1) and is thus equal to $(d/dA)e^{(1/2)A^2} = Ae^{(1/2)A^2}$. Therefore the expression (9.1) is equal to

$$\frac{A}{\sqrt{1/v^2 + m/\sigma^2}} = \frac{mz}{\sigma^2/v^2 + m}$$

Thus the decision chosen by the decision rule s_v is

$$\frac{mZ}{\sigma^2/v^2 + m} \tag{9.2}$$

$$r(\theta;s_v) = E\left\{\left(\frac{mZ}{m + \sigma^2/v^2} - \theta\right)^2\right\} = E\left\{\left[\frac{m(Z-\theta) - (\sigma^2/v^2)\theta}{m + \sigma^2/v^2}\right]^2\right\}$$

$$= \frac{m^2}{(m + \sigma^2/v^2)^2}\,E\{(Z-\theta)^2\} + \frac{\theta^2}{(1 + mv^2/\sigma^2)^2}$$

$$= \frac{mv^4/\sigma^2 + \theta^2}{(1 + mv^2/\sigma^2)^2}$$

9.6. An Admissible Decision Rule Which Is a Limit of Bayes Decision Rules. Continuing our discussion of the example of Sec. 9.5, we note that

$$\lim_{v\to\infty} r(\theta;s_v) = \frac{\sigma^2}{m}$$

Denoting by s the decision rule which chooses the decision Z, it is easily verified that $r(\theta;s) = \sigma^2/m$. Therefore the decision rule s is the limit of the decision rules s_v as v increases, in the sense of Sec. 5.10. We are going to show that s is a minimax and admissible decision rule.

First we note that if s is admissible, it must be minimax. For if s is not minimax, there would be a decision rule t with

$$\max_\theta r(\theta;t) < \max_\theta r(\theta;s) = \frac{\sigma^2}{m}$$

and therefore $r(\theta;t) < \sigma^2/m = r(\theta;s)$ for all θ, so that s would not be admissible. Therefore we have only to show that s is admissible, and this will also show that s is minimax.

In order to show that s is admissible, we assume that it is not admissible and force a contradiction. If s is not admissible, there is a decision rule t with $r(\theta;t) \leqslant r(\theta;s)$ for all θ, and $r(\theta;t) < r(\theta;s)$ for at least one value of θ. But since $r(\theta;t)$ and $r(\theta;s)$ are both continuous functions of θ in the present problem, there must be values A, B, Δ, with $B > A$ and $\Delta > 0$, such that $r(\theta;t) \leqslant r(\theta;s) - \Delta$ for all θ between A and B. Then we have, for each $v > 0$,

$$v\int_{-\infty}^{\infty}[r(\theta;s) - r(\theta;t)]b_v(\theta)\,d\theta \geqslant v\int_A^B [r(\theta;s) - r(\theta;t)]b_v(\theta)\,d\theta$$

$$\geqslant v\int_A^B \Delta\frac{1}{v\sqrt{2\pi}}\,e^{-(\theta^2/2v^2)}\,d\theta = \frac{\Delta}{\sqrt{2\pi}}\int_A^B e^{-(\theta^2/2v^2)}\,d\theta$$

and this last expression approaches $\Delta(B-A)/\sqrt{2\pi}$ as v increases.

Therefore, as v increases, $v\int_{-\infty}^{\infty} [r(\theta;s) - r(\theta;t)]b_v(\theta)\, d\theta$ remains no smaller than a quantity very close to $\Delta(B - A)/\sqrt{2\pi}$, which is positive. It is easily verified that

$$v\int_{-\infty}^{\infty} [r(\theta;s_v) - r(\theta;s)]b_v(\theta)\, d\theta = \frac{-v}{1 + mv^2/\sigma^2}$$

which approaches zero as v increases. But

$$v\int_{-\infty}^{\infty} [r(\theta;s_v) - r(\theta;t)]b_v(\theta)\, d\theta = v\int_{-\infty}^{\infty} [r(\theta;s_v) - r(\theta;s)]b_v(\theta)\, d\theta + v\int_{-\infty}^{\infty} [r(\theta;s) - r(\theta;t)]b_v(\theta)\, d\theta$$

and therefore as v increases, $v\int_{-\infty}^{\infty} [r(\theta;s_v) - r(\theta;t)]b_v(\theta)\, d\theta$ becomes positive. This means that a value v' can be found such that $v'\int_{-\infty}^{\infty} [r(\theta;s_{v'}) - r(\theta;t)]b_{v'}(\theta)\, d\theta$ is positive, which implies that

$$\int_{-\infty}^{\infty} r(\theta;s_{v'})b_{v'}(\theta)\, d\theta > \int_{-\infty}^{\infty} r(\theta;t)b_{v'}(\theta)\, d\theta$$

But this last inequality is a contradiction, by the definition of $s_{v'}$ as a Bayes decision rule relative to $B_{v'}(\theta)$. This contradiction proves that s is admissible.

9.7. Estimation of Location Parameters. If $f(x_1, \dots, x_m; \theta)$ can be written in the form $g(x_1 - \theta, x_2 - \theta, \dots, x_m - \theta)$, then θ is called a "location parameter." The reason is as follows. If we define the chance variables $Y_1, \dots, Y_m$ by $Y_1 = X_1 + c$, $Y_2 = X_2 + c, \dots, Y_m = X_m + c$, where c is an arbitrary constant, then it is easily seen that the joint pdf for $Y_1, \dots, Y_m$ is $g(y_1 - \theta - c, y_2 - \theta - c, \dots, y_m - \theta - c)$. But this is the same as the pdf for $X_1, \dots, X_m$, except that θ has been increased by c. Thus the addition of a constant to each variable (that is, a "change of location") has the effect of adding the same constant to θ in the joint pdf.

Suppose the problem is to estimate θ, and suppose that $W(\theta;D) = (D - \theta)^2$. If $k(X_1, \dots, X_m)$ denotes the decision chosen by our decision rule when $X_1, \dots, X_m$ are observed, it seems reasonable, because of the discussion in the preceding paragraph, to demand that $k(X_1 + c, \dots, X_m + c) = k(X_1, \dots, X_m) + c$, for all values of $X_1, \dots, X_m$ and c. For the remainder of this section, we are going to limit attention to decision rules which satisfy this condition, which is called an "invariance condition."

If $k(X_1, \ldots, X_m)$ is the decision chosen by the decision rule s and $k(X_1 + c, \ldots, X_m + c) = k(X_1, \ldots, X_m) + c$ for all values $X_1, \ldots, X_m$ and c, then

$$r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [k(x_1, \ldots, x_m) - \theta]^2 g(x_1 - \theta, \ldots, x_m - \theta)\, dx_1 \cdots dx_m$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [k(x_1 - \theta, \ldots, x_m - \theta)]^2 \times g(x_1 - \theta, \ldots, x_m - \theta)\, dx_1 \cdots dx_m$$

In this last integral, if we make the change of variables $y_1 = x_1 - \theta$, $y_2 = x_2 - \theta, \ldots, y_m = x_m - \theta$, we get

$$r(\theta;s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [k(y_1, \ldots, y_m)]^2 g(y_1, \ldots, y_m)\, dy_1 \cdots dy_m$$

and since this integral does not depend on θ, we see that $r(\theta;s)$ does not depend on θ, if s satisfies the invariance condition. Clearly, we should like to choose $k(y_1, \ldots, y_m)$ to minimize the integral

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [k(y_1, \ldots, y_m)]^2 g(y_1, \ldots, y_m)\, dy_1 \cdots dy_m$$

By the invariance condition, $k(y_1 - y_m, \ldots, y_{m-1} - y_m, y_m - y_m) = k(y_1, \ldots, y_{m-1}, y_m) - y_m$. Therefore we have $k(y_1, \ldots, y_{m-1}, y_m) = y_m + k(y_1 - y_m, \ldots, y_{m-1} - y_m, 0)$. Putting this expression into the integral to be minimized, we get

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [y_m + k(y_1 - y_m, \ldots, y_{m-1} - y_m, 0)]^2 g(y_1, \ldots, y_m)\, dy_1 \cdots dy_m$$

In this integral, we change to the variables $t_1, \ldots, t_m$ defined as follows: $t_1 = y_1 - y_m$, $t_2 = y_2 - y_m, \ldots, t_{m-1} = y_{m-1} - y_m$, $t_m = y_m$. Then the integral to be minimized becomes

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [t_m + k(t_1, \ldots, t_{m-1}, 0)]^2 g(t_1 + t_m, \ldots, t_{m-1} + t_m, t_m)\, dt_1 \cdots dt_m$$

It is easily seen that this last integral will be minimized if, for each given set of values $t_1, \ldots, t_{m-1}$, we set the value of $k(t_1, \ldots, t_{m-1}, 0)$ so as to minimize

$$\int_{-\infty}^{\infty} [t_m + k(t_1, \ldots, t_{m-1}, 0)]^2 g(t_1 + t_m, \ldots, t_{m-1} + t_m, t_m)\, dt_m$$

Differentiating with respect to $k(t_1, t_2, \ldots, t_{m-1}, 0)$ and setting the result equal to zero, we find that the minimizing value of $k(t_1, \ldots, t_{m-1}, 0)$ is

$$-\frac{\int_{-\infty}^{\infty} t_m g(t_1 + t_m, \ldots, t_{m-1} + t_m, t_m)\, dt_m}{\int_{-\infty}^{\infty} g(t_1 + t_m, \ldots, t_{m-1} + t_m, t_m)\, dt_m} \tag{9.3}$$

In any particular problem, (9.3) is computed, and then if $x_1, \ldots, x_m$ are the observed values, the decision chosen by the best decision rule satisfying the invariance condition is $x_m + k(x_1 - x_m, \ldots, x_{m-1} - x_m, 0)$.

Before discussing examples, we note that in the special case when $m = 1$, the expression (9.3) becomes

$$-\frac{\int_{-\infty}^{\infty} tg(t)\, dt}{\int_{-\infty}^{\infty} g(t)\, dt} \tag{9.4}$$

As a first example, suppose $f(x_1, \ldots, x_m; \theta)$ is given by

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i - \theta)^2\right]$$

where σ is known. Defining Z as $(1/m)(X_1 + \cdots + X_m)$, we know that Z is sufficient for the decision problem. The pdf for Z is

$$\frac{\sqrt{m}}{\sigma\sqrt{2\pi}} \exp\left[-\frac{m}{2\sigma^2}(z - \theta)^2\right]$$

and we note that θ remains a location parameter in the distribution for Z. Since we are basing our decision on the single variable Z, we use formula (9.4) above, getting

$$-\frac{\int_{-\infty}^{\infty} t \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{m}{2\sigma^2}t^2\right) dt}{\int_{-\infty}^{\infty} \frac{\sqrt{m}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{m}{2\sigma^2}t^2\right) dt} = 0$$

Therefore the best decision rule satisfying the invariance condition is to choose the decision z when the observed value of Z is z. This decision rule has already been shown to be minimax and admissible, in Sec. 9.6.

As another example, suppose $X_1, \ldots, X_m$ are all independent, each with a uniform distribution between $\theta - \frac{1}{2}$ and $\theta + \frac{1}{2}$. Then $f(x_1, \ldots, x_m; \theta) = 1$ if $-\frac{1}{2} < x_1 - \theta < \frac{1}{2}$, $-\frac{1}{2} < x_2 - \theta < \frac{1}{2}, \ldots,$ and $-\frac{1}{2} < x_m - \theta < \frac{1}{2}$, and $f(x_1, \ldots, x_m; \theta) = 0$ if $|x_i - \theta| > \frac{1}{2}$

for any i. Thus in computing (9.3), the integrands are zero unless all the following inequalities hold:

$$\begin{gathered}-\tfrac{1}{2} < t_1 + t_m < \tfrac{1}{2}\\ -\tfrac{1}{2} < t_2 + t_m < \tfrac{1}{2}\\ \vdots\\ -\tfrac{1}{2} < t_{m-1} + t_m < \tfrac{1}{2}\\ -\tfrac{1}{2} < t_m < \tfrac{1}{2}\end{gathered}$$

Denoting by L_1 the largest of the quantities $-\frac{1}{2} - t_1, -\frac{1}{2} - t_2, \ldots, -\frac{1}{2} - t_{m-1}, -\frac{1}{2}$, and by L_2 the smallest of the quantities $\frac{1}{2} - t_1, \frac{1}{2} - t_2, \ldots, \frac{1}{2} - t_{m-1}, \frac{1}{2}$, we see that the inequalities above are equivalent to $L_1 < t_m < L_2$. Furthermore, when $L_1 < t_m < L_2$, $g(t_1 + t_m, \ldots, t_{m-1} + t_m, t_m) = 1$. Therefore in the present problem (9.3) becomes

$$-\frac{\int_{L_1}^{L_2} t_m\,dt_m}{\int_{L_1}^{L_2} dt_m} = -(\tfrac{1}{2})(L_2 + L_1)$$

and the decision chosen by the best decision rule satisfying the invariance condition is

$$x_m - (\tfrac{1}{2})[\min(\tfrac{1}{2} - x_1 + x_m, \tfrac{1}{2} - x_2 + x_m, \ldots, \tfrac{1}{2} - x_{m-1} + x_m, \tfrac{1}{2}) + \max(-\tfrac{1}{2} - x_1 + x_m, -\tfrac{1}{2} - x_2 + x_m, \ldots, -\tfrac{1}{2} - x_{m-1} + x_m, -\tfrac{1}{2})]$$

Since

$$\min(\tfrac{1}{2} - x_1 + x_m, \tfrac{1}{2} - x_2 + x_m, \ldots, \tfrac{1}{2} - x_{m-1} + x_m, \tfrac{1}{2})$$

is equal to

$$\tfrac{1}{2} + x_m - \max(x_1, x_2, \ldots, x_m)$$

and

$$\max(-\tfrac{1}{2} - x_1 + x_m, -\tfrac{1}{2} - x_2 + x_m, \ldots, -\tfrac{1}{2} - x_{m-1} + x_m, -\tfrac{1}{2})$$

is equal to

$$-\tfrac{1}{2} + x_m - \min(x_1, x_2, \ldots, x_m)$$

the decision chosen is

$$x_m - (\tfrac{1}{2})[\tfrac{1}{2} + x_m - \max(x_1, \ldots, x_m) - \tfrac{1}{2} + x_m - \min(x_1, \ldots, x_m)]$$

which is equal to

$$(\tfrac{1}{2})[\max(x_1, \ldots, x_m) + \min(x_1, \ldots, x_m)]$$

In this section we have limited our attention to decision rules satisfying a given invariance condition and have found the best such decision rule. But what guarantee is there that this best invariant rule is admissible when compared with all possible decision rules? In the first of our two examples, we knew from other considerations that the rule was admissible. And in most standard cases, a development like that in Sec. 9.6 can be used to show that the best invariant rule is admissible. If an invariant rule is admissible, it is also minimax: this follows from the same reasoning as that used in the second paragraph of Sec. 9.6.

9.8. Estimation of Scale Parameters. If $X_1, \ldots, X_m$ are chance variables which are nonnegative with probability 1, and if $f(x_1, \ldots, x_m; \theta)$ can be written in the form $(1/\theta^m)g(x_1/\theta, \ldots, x_m/\theta)$, where θ is positive, then θ is called a "scale parameter." The reason is as follows. If we define the chance variables $Y_1, \ldots, Y_m$ by $Y_1 = cX_1$, $Y_2 = cX_2, \ldots$, $Y_m = cX_m$, where c is an arbitrary positive constant, then it is easily seen that the joint pdf for $Y_1, \ldots, Y_m$ is $[1/(c\theta)^m]g(y_1/c\theta, \ldots, y_m/c\theta)$. But this is the same as the pdf for $X_1, \ldots, X_m$, except that θ has been multiplied by c. Thus the multiplication of each variable by a positive constant (that is, a "change of scale") has the effect of multiplying θ by the same constant in the joint pdf.

Suppose the problem is to estimate θ, and suppose that $W(\theta;D) = (1/\theta^2)(D - \theta)^2$. If $k(X_1, \ldots, X_m)$ denotes the decision chosen by our decision rule when $X_1, \ldots, X_m$ are observed, it seems reasonable, because of the discussion in the preceding paragraph, to demand that $k(cX_1, \ldots, cX_m) = ck(X_1, \ldots, X_m)$, for all positive values $X_1, \ldots, X_m$ and c. For the remainder of this section, we are going to limit attention to decision rules which satisfy this invariance condition.

If $k(X_1, \ldots, X_m)$ is the decision chosen by a decision rule s satisfying the invariance condition, then

$$r(\theta;s) = \int_0^\infty \cdots \int_0^\infty \left[\frac{k(x_1, \ldots, x_m) - \theta}{\theta}\right]^2 \frac{1}{\theta^m} g\left(\frac{x_1}{\theta}, \ldots, \frac{x_m}{\theta}\right) dx_1 \cdots dx_m$$

$$= \int_0^\infty \cdots \int_0^\infty \left[k\left(\frac{x_1}{\theta}, \ldots, \frac{x_m}{\theta}\right) - 1\right]^2 g\left(\frac{x_1}{\theta}, \ldots, \frac{x_m}{\theta}\right) d\left(\frac{x_1}{\theta}\right) \cdots d\left(\frac{x_m}{\theta}\right)$$

In this last integral, if we make the change of variables $y_1 = x_1/\theta, \ldots$, $y_m = x_m/\theta$, we get

$$r(\theta;s) = \int_0^\infty \cdots \int_0^\infty [k(y_1, \ldots, y_m) - 1]^2 g(y_1, \ldots, y_m)\, dy_1 \cdots dy_m$$

and since this integral does not depend on θ, $r(\theta;s)$ does not depend on θ. Clearly, we should like to choose $k(y_1, \ldots, y_m)$ to minimize the integral

$\int_0^\infty \cdots \int_0^\infty [k(y_1, \ldots, y_m) - 1]^2 g(y_1, \ldots, y_m)\, dy_1 \cdots dy_m$. By the invariance condition, $k(y_1/y_m, y_2/y_m, \ldots, y_{m-1}/y_m, y_m/y_m) = (1/y_m)k(y_1, \ldots, y_{m-1}, y_m)$. Therefore we have $k(y_1, \ldots, y_{m-1}, y_m) = y_m k(y_1/y_m, \ldots, y_{m-1}/y_m, 1)$. Putting this expression into the integral to be minimized, we get

$$\int_0^\infty \cdots \int_0^\infty \left[y_m k\left(\frac{y_1}{y_m}, \ldots, \frac{y_{m-1}}{y_m}, 1\right) - 1 \right]^2 g(y_1, \ldots, y_m)\, dy_1 \cdots dy_m$$

In this integral, we change to the variables $t_1, \ldots, t_m$ defined as follows: $t_1 = y_1/y_m, \ldots, t_{m-1} = y_{m-1}/y_m, t_m = y_m$. Then the integral to be minimized becomes

$$\int_0^\infty \cdots \int_0^\infty [t_m k(t_1, \ldots, t_{m-1}, 1) - 1]^2 t_m^{m-1} g(t_1 t_m, \ldots, t_{m-1} t_m, t_m)\, dt_1 \cdots dt_m$$

It is easily seen that this last integral will be minimized if, for each given set of values $t_1, \ldots, t_{m-1}$, we set the value of $k(t_1, \ldots, t_{m-1}, 1)$ so as to minimize

$$\int_0^\infty [t_m k(t_1, \ldots, t_{m-1}, 1) - 1]^2 t_m^{m-1} g(t_1 t_m, \ldots, t_{m-1} t_m, t_m)\, dt_m$$

Differentiating with respect to $k(t_1, \ldots, t_{m-1}, 1)$ and setting the result equal to zero, we find that the minimizing value of $k(t_1, \ldots, t_{m-1}, 1)$ is

$$\frac{\int_0^\infty t_m{}^m g(t_1 t_m, \ldots, t_{m-1} t_m, t_m)\, dt_m}{\int_0^\infty t_m^{m+1} g(t_1 t_m, \ldots, t_{m-1} t_m, t_m)\, dt_m} \tag{9.5}$$

In any particular problem, (9.5) is computed, and then if $x_1, \ldots, x_m$ are the observed values, the decision chosen by the best decision rule satisfying the invariance condition is $x_m k(x_1/x_m, \ldots, x_{m-1}/x_m, 1)$.

Before discussing examples, we note that in the special case when $m = 1$, the expression (9.5) becomes

$$\frac{\int_0^\infty t g(t)\, dt}{\int_0^\infty t^2 g(t)\, dt} \tag{9.6}$$

Our first example will violate one of our assumptions, since the variables $X_1, \ldots, X_m$ will not have to be nonnegative. However, there will be a chance variable Z sufficient for the decision problem, and Z will have to be nonnegative, so in terms of Z our assumptions will be satisfied.

Suppose $X_1, \ldots, X_m$ are independent, each with a normal distribution with the same parameters: known mean μ and unknown standard deviation θ. Thus

$$f(x_1, \ldots, x_m; \theta) = \left(\frac{1}{\theta\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{m}(x_i - \mu)^2\right]$$

Then it is easily verified that $Z = \sqrt{\sum_{1}^{m}(X_i - \mu)^2}$ is sufficient for the decision problem. From Sec. 4.8, we know that Z^2/θ^2 has a chi-square distribution with m degrees of freedom. From this, using the technique described in Sec. 3.7, we find that the pdf for Z, $g(z;\theta)$, say, is given as follows:

$$g(z;\theta) = 0 \qquad \text{if } z < 0$$

$$g(z;\theta) = \frac{1}{2^{(m-2)/2}\Gamma(m/2)}\,\frac{1}{\theta}\left(\frac{z}{\theta}\right)^{m-1} e^{-(\frac{1}{2})(z/\theta)^2} \qquad \text{if } z > 0$$

We note that θ is a scale parameter in the distribution for Z. Since we are basing our decision on the single variable Z, we use formula (9.6) above, getting (after canceling common constant factors)

$$\frac{\int_0^\infty t(t^{m-1}e^{-(\frac{1}{2})t^2})\,dt}{\int_0^\infty t^2(t^{m-1}e^{-(\frac{1}{2})t^2})\,dt} \tag{9.7}$$

To evaluate (9.7), we note that we can evaluate $\int_0^\infty t^r e^{-(\frac{1}{2})t^2}\,dt$ by making the change of variable $w = (\frac{1}{2})t^2$, the integral then becoming

$$2^{(r-1)/2}\int_0^\infty w^{(r+1)/2-1}e^{-w}\,dw = 2^{(r-1)/2}\Gamma\left(\frac{r+1}{2}\right)$$

Thus (9.7) is equal to

$$\frac{2^{(m-1)/2}\Gamma[(m+1)/2]}{2^{m/2}\Gamma[(m+2)/2]} = \frac{1}{\sqrt{2}}\,\frac{\Gamma[(m+1)/2]}{\Gamma[(m+2)/2]}$$

and the best invariant decision rule is to choose the decision

$$z\left\{\frac{1}{\sqrt{2}}\,\frac{\Gamma[(m+1)/2]}{\Gamma[(m+2)/2]}\right\}$$

when the observed value of Z is z.

As a second example, suppose $X_1, \ldots, X_m$ are all independent, each with a uniform distribution between 0 and θ. Thus $f(x_1, \ldots, x_m; \theta) = 1/\theta^m$ if $0 < x_1/\theta < 1$, $0 < x_2/\theta < 1, \ldots,$ and $0 < x_m/\theta < 1$, and $f(x_1, \ldots, x_m; \theta) = 0$ if any of the quantities $x_1, \ldots, x_m$ is below zero or above θ. We note that in this problem, the function $g(y_1, \ldots, y_m)$ is equal to 1 if all the quantities $y_1, \ldots, y_m$ are between 0 and 1 and is equal to 0 if any of the quantities $y_1, \ldots, y_m$ is below 0 or above 1. Then in computing the expression (9.5) for this case, we see that the integrands are 0 unless the following inequalities all hold:

$$\begin{gathered} t_1 t_m < 1 \\ t_2 t_m < 1 \\ \cdot \\ \cdot \\ \cdot \\ t_{m-1} t_m < 1 \\ t_m < 1 \end{gathered}$$

Letting U denote $\min(1/t_1, 1/t_2, \ldots, 1/t_{m-1}, 1)$, the inequalities are equivalent to $t_m < U$. Therefore (9.5) becomes

$$\frac{\int_0^U t_m{}^m 1 \, dt_m}{\int_0^U t_m^{m+1} 1 \, dt_m} = \frac{m+2}{m+1} \frac{1}{U}$$

Thus the decision chosen by the best invariant decision rule is

$$x_m \frac{m+2}{m+1} \frac{1}{U'}$$

where $U' = \min(x_m/x_1, x_m/x_2, \ldots, x_m/x_{m-1}, 1)$. But we can write

$$U' = \min\left(\frac{x_m}{x_1}, \frac{x_m}{x_2}, \ldots, \frac{x_m}{x_{m-1}}, \frac{x_m}{x_m}\right) = x_m \min\left(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_{m-1}}, \frac{1}{x_m}\right)$$
$$= \frac{x_m}{\max(x_1, x_2, \ldots, x_{m-1}, x_m)}$$

so that $1/U' = (1/x_m) \max(x_1, x_2, \ldots, x_{m-1}, x_m)$. Therefore the decision chosen by the best invariant decision rule is

$$\frac{m+2}{m+1} \max(x_1, \ldots, x_m)$$

9.9. Interval Estimation. In the preceding sections, we have been discussing the problem of point estimation. Another type of estimation is "interval estimation," in which the possible decisions are intervals. We should like the interval we choose to contain the true value of θ, and we should also like the interval to be short in length. Let D_1 denote the lower end point of the interval we choose and D_2 denote the upper end point of the interval. Then a reasonable type of loss function would seem to be as follows:

$$\begin{aligned} W(\theta;D_1,D_2) &= c(\theta)(D_2 - D_1) \qquad \text{if } D_1 < \theta \leqslant D_2 \\ &= A + c(\theta)(D_2 - D_1) \qquad \text{if } \theta < D_1 \text{ or } \theta > D_2 \end{aligned}$$

where A = a given positive constant
$c(\theta)$ = a given function of θ which is never negative

With the type of loss function introduced in the preceding paragraph, suppose we want to construct a Bayes decision rule relative to an a priori distribution $B(\theta)$ with pdf $b(\theta)$. Then

$$\begin{aligned} K(D_1,D_2;x) = &\int_{\theta< D_1} [A + c(\theta)(D_2 - D_1)]b(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta \\ &+ \int_{D_1}^{D_2} c(\theta)(D_2 - D_1)b(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta \\ &+ \int_{\theta>D_2} [A + c(\theta)(D_2 - D_1)]b(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta \end{aligned}$$

The Bayes decision rule chooses the values of D_1, D_2 (with $D_1 \leqslant D_2$) which minimize $K(D_1,D_2;x)$. The equation $(\partial/\partial D_1)K(D_1,D_2;x) = 0$ gives

$$Ab(D_1)f(x_1, \ldots, x_m; D_1) = \int c(\theta)b(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta \qquad (9.8)$$

and the equation $(\partial/\partial D_2)K(D_1,D_2;x) = 0$ gives

$$Ab(D_2)f(x_1, \ldots, x_m; D_2) = \int c(\theta)b(\theta)f(x_1, \ldots, x_m; \theta)\, d\theta \qquad (9.9)$$

To ensure a minimum, in any specific example the second derivatives of $K(D_1,D_2;x)$ should be examined.

As an example, suppose $f(x_1, \ldots, x_m; \theta)$ is

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i - \theta)^2\right]$$

where σ = a known quantity
$c(\theta) = C$, a positive constant

Defining Z as $(1/m)(X_1 + \cdots + X_m)$, it is known that Z is sufficient for the decision problem. The pdf for Z, $g(z;\theta)$, say, is

$$(\sqrt{m}/\sigma\sqrt{2\pi}) \exp\,[-(m/2\sigma^2)(z - \theta)^2]$$

Suppose we want to construct a Bayes decision rule s_v relative to the a priori distribution with

$$\text{pdf } b_v(\theta) = (1/v\sqrt{2\pi}) \exp\,[-(\theta^2/2v^2)]$$

Then equations (9.8)and (9.9) become

$$A\,\frac{1}{v\sqrt{2\pi}}\exp\left(-\frac{D_1{}^2}{2v^2}\right)\frac{\sqrt{m}}{\sigma\sqrt{2\pi}}\exp\left[-\frac{m}{2\sigma^2}(z-D_1)^2\right]$$
$$=C\int_{-\infty}^{\infty}\frac{1}{v\sqrt{2\pi}}\exp\left(-\frac{\theta^2}{2v^2}\right)\frac{\sqrt{m}}{\sigma\sqrt{2\pi}}\exp\left[-\frac{m}{2\sigma^2}(z-\theta)^2\right]d\theta$$

$$A\,\frac{1}{v\sqrt{2\pi}}\exp\left(-\frac{D_2{}^2}{2v^2}\right)\frac{\sqrt{m}}{\sigma\sqrt{2\pi}}\exp\left[-\frac{m}{2\sigma^2}(z-D_2)^2\right]$$
$$=C\int_{-\infty}^{\infty}\frac{1}{v\sqrt{2\pi}}\exp\left(-\frac{\theta^2}{2v^2}\right)\frac{\sqrt{m}}{\sigma\sqrt{2\pi}}\exp\left[-\frac{m}{2\sigma^2}(z-\theta)^2\right]d\theta$$

The integral appearing in these equations has been evaluated in Sec. 9.5: its value is

$$\frac{R\sqrt{m}}{\sigma v\sqrt{2\pi}}\exp\left(-\frac{mz^2}{2\sigma^2}\right)\exp\left(\frac{1}{2}\frac{R^2m^2z^2}{\sigma^4}\right)$$

where $R = (1/v^2 - m/\sigma^2)^{-1/2}$. Simplifying the equations and taking logs, we find that both D_1 and D_2 satisfy the following quadratic equation in D:

$$-\frac{D^2}{2R^2}+\frac{mz}{\sigma^2}D=\log\frac{CR}{A}+\frac{1}{2}\frac{R^2m^2z^2}{\sigma^4} \tag{9.10}$$

If $CR < A$, (9.10) has two distinct real roots, and D_1 is the lower of the two roots, D_2 is the higher of the two roots. If $CR \geq A$, (9.10) does not have two distinct real roots, and we have a degenerate situation where $D_1 = D_2$.

9.10. Estimation by the Method of Maximum Likelihood. Suppose that the possible joint distributions of $X_1, \ldots, X_m$ are given by the variation of u parameters, say, $\theta_1, \ldots, \theta_u$, and the problem is to construct a point estimate for each parameter. Let D_i denote the estimate of θ_i. The following decision rule, known as the method of "maximum likelihood," is currently in very wide use: Choose the values of D_1, $D_2, \ldots, D_u$ so that

$$f(x_1, \ldots, x_m; D_1, D_2, \ldots, D_u) = \max_{\theta_1,\theta_2,\ldots,\theta_u} f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_u)$$

We give several examples of the maximum-likelihood decision rule.

Example 1. $X_1, X_2, \ldots, X_m$ are all independent, each with a normal distribution with known standard deviation σ and unknown mean θ. The maximum-likelihood decision rule chooses the value of θ that maximizes

$$f(x_1, \ldots, x_m; \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i - \theta)^2\right]$$

This is the same value that maximizes

$$\log f(x_1, \ldots, x_m; \theta) = -m \log \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i - \theta)^2$$

$$\frac{\partial}{\partial\theta}\log f(x_1, \ldots, x_m; \theta) = \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i - \theta)$$

and solving the equation

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i - \theta) = 0$$

we get

$$\sum_{i=1}^{m} x_i - m\theta = 0 \qquad \text{or} \qquad \theta = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Thus the maximum-likelihood decision rule is to choose the decision $(1/m)(x_1 + \cdots + x_m)$ when the observed values are $x_1, \ldots, x_m$.

Example 2. $X_1, \ldots, X_m$ are all independent, each with a normal distribution with known mean μ and unknown standard deviation θ. Then

$$f(x_1, \ldots, x_m; \theta) = \left(\frac{1}{\theta\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{m}(x_i - \mu)^2\right]$$

and

$$\frac{\partial}{\partial\theta}\log f(x_1, \ldots, x_m; \theta) = -\frac{m}{\theta} + \frac{\sum(x_i - \mu)^2}{\theta^3}$$

Solving the equation

$$-\frac{m}{\theta} + \frac{\sum(x_i - \mu)^2}{\theta^3} = 0$$

we get $\theta = \sqrt{(1/m)\,\Sigma(x_i - \mu)^2}$. Thus, the maximum-likelihood decision rule is to choose the decision $\sqrt{(1/m)\,\Sigma(x_i - \mu)^2}$ when the observed values are $x_1, \ldots, x_m$.

Example 3. $X_1, \ldots, X_m$ are all independent, each with a normal distribution with unknown mean θ_1 and unknown standard deviation θ_2. Both are to be estimated.

$$f(x_1, \ldots, x_m; \theta_1, \theta_2) = \left(\frac{1}{\theta_2\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\theta_2^{\,2}}\sum_{i=1}^{m}(x_i - \theta_1)^2\right]$$

The equations

$$\frac{\partial \log f(x_1, \ldots, x_m; \theta_1, \theta_2)}{\partial\theta_1} = 0 \qquad \text{and} \qquad \frac{\partial \log f(x_1, \ldots, x_m; \theta_1, \theta_2)}{\partial\theta_2} = 0$$

give

$$\frac{1}{\theta_2^{\,2}}\sum_{i=1}^{m}(x_i - \theta_1) = 0$$

$$-\frac{m}{\theta_2} + \frac{\sum(x_i - \theta_1)^2}{\theta_2^{\,3}} = 0$$

The first equation gives $(1/m)(x_1 + \cdots + x_m)$ as the maximizing value of θ_1. Denoting $(1/m)(x_1 + \cdots + x_m)$ by $\bar{x}$ for convenience, the second equation gives $\sqrt{(1/m) \sum_1^m (x_i - \bar{x})^2}$ as the maximizing value of θ_2.

Example 4. $X_1, \ldots, X_m$ are all independent, each with a uniform distribution over $\theta - \frac{1}{2}$, $\theta + \frac{1}{2}$. Thus $f(x_1, \ldots, x_m; \theta) = 1$ if $\theta - \frac{1}{2} \leq x_1 \leq \theta + \frac{1}{2}$, $\theta - \frac{1}{2} \leq x_2 \leq \theta + \frac{1}{2}, \ldots,$ and $\theta - \frac{1}{2} \leq x_m \leq \theta + \frac{1}{2}$, and $f(x_1, \ldots, x_m; \theta) = 0$ for other values of $x_1, \ldots, x_m$. Let L_1 denote min $(x_1, \ldots, x_m)$ and L_2 denote max $(x_1, \ldots, x_m)$. Then, for any value of θ between $L_2 - \frac{1}{2}$ and $L_1 + \frac{1}{2}$, the value of $f(x_1, \ldots, x_m; \theta)$ is equal to 1. Thus, in this case, there is no unique maximizing value of θ.

Example 5. $X_1, \ldots, X_m$ are all independent, each with a uniform distribution over 0, θ, where θ is a positive quantity which is otherwise unknown. Then $f(x_1, \ldots, x_m; \theta) = 1/\theta^m$ for $0 \leq x_1 \leq \theta$, $0 \leq x_2 \leq \theta, \ldots,$ and $0 \leq x_m \leq \theta$ and is equal to zero for other values of $x_1, \ldots, x_m$. Let L denote max $(x_1, \ldots, x_m)$. Then $f(x_1, \ldots, x_m; \theta) = 1/\theta^m$ for $\theta \geq L$, and $f(x_1, \ldots, x_m; \theta) = 0$ for $\theta < L$. Clearly, the maximizing value of θ is L, and so the maximum-likelihood decision rule is to set the decision equal to the largest of the observed values $x_1, \ldots, x_m$.

We have stated that the maximum-likelihood decision rule is very widely used. Since this is the case, it might be thought that the maximum-likelihood decision rule is a good decision rule in the sense that the expected loss function $r(\theta; s)$ is small when s is the maximum-likelihood decision rule. However, this is not always true. Actually, it is known that if $X_1, \ldots, X_m$ are all independent with the same distributions (as is the case in our examples above) and if the loss function is of the form $c(\theta)(D - \theta)^2$, then if certain restrictions are satisfied by the distribution of the X's, the maximum-likelihood decision rule comes closer and closer to being a minimax decision rule as m increases. We shall not prove this statement in this text, but we point out that what happens as m approaches infinity is of limited help in practical problems, where m is fixed.

Our examples above give some information about how good the maximum-likelihood decision rule is. In Example 1, the maximum-likelihood decision rule has been shown to be admissible and minimax in Sec. 9.6. In Example 2, if we denote $\sqrt{\sum_1^m (X_i - \mu)^2}$ by Z, we know from Sec. 9.8 that θ is a scale parameter in the distribution of Z. The maximum-likelihood decision rule in Example 2 satisfies the invariance condition of Sec. 9.8, but is *not* the same as the best invariant decision rule found in Sec. 9.8. Thus the maximum-likelihood decision rule in Example 2 is inadmissible. Exactly the same conclusion holds in

Example 5, for the same reasons. Thus the maximum-likelihood decision rule is sometimes a poor decision rule.

9.11. Testing a Hypothesis by the Likelihood Ratio Method. The following decision rule s, known as the "likelihood-ratio rule, "is currently in very wide use for the problem of testing a hypothesis: Choose $D = 1$ if

$$\frac{\max\limits_{\theta \text{ in I}} f(x_1, \ldots, x_m; \theta)}{\max\limits_{\text{II}} f(x_1, \ldots, x_m; \theta)} > k$$

where k is a constant chosen so that

$$\max_{\theta \text{ in I}} r(\theta; s) = \alpha$$

The quantity

$$\frac{\max\limits_{\theta \text{ in I}} f(x_1, \ldots, x_m; \theta)}{\max\limits_{\theta} f(x_1, \ldots, x_m; \theta)}$$

is known as the "likelihood ratio."

We give some examples of the use of the likelihood-ratio rule.

Example 1. X_1, X_2, X_3, X_4 are independent, each with a normal distribution with standard deviation equal to 1 and unknown mean θ, and group I consists of all distributions given by values of θ less than or equal to 2. (Note that it is not necessary to specify group II in order to use the likelihood-ratio decision rule.) $\alpha = 0.1$. Denoting $(\frac{1}{4})(x_1 + x_2 + x_3 + x_4)$ by z, we know from Example 1 of Sec. 9.10 that

$$\max_{\theta} f(x_1, \ldots, x_4; \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^4 \exp\left[-\frac{1}{2}\sum_{i=1}^{4}(x_i - z)^2\right]$$

It is easily seen that

$$\max_{\theta \text{ in I}} f(x_1, \ldots, x_4; \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^4 \exp\left[-\frac{1}{2}\sum_{i=1}^{4}(x_i - z)^2\right] \qquad \text{if } z < 2$$

$$\max_{\theta \text{ in I}} f(x_1, \ldots, x_4; \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^4 \exp\left[-\frac{1}{2}\sum_{1}^{4}(x_i - 2)^2\right] \qquad \text{if } z > 2$$

Therefore the likelihood ratio is equal to 1 if $z \leq 2$ and is equal to

$$\exp\left[-\frac{1}{2}\sum_{1}^{4}(x_i - 2)^2 + \frac{1}{2}\sum_{1}^{4}(x_i - z)^2\right] = \exp\left[-2(z-2)^2\right] \qquad \text{if } z > 2$$

Since $\exp\left[-2(z-2)^2\right]$ decreases as z increases above 2, it is clear that the likelihood-ratio test in this case chooses $D = 1$ when $z \leq c$, where c is a value chosen so that

$$\max_{\theta \text{ in I}} r(\theta; s) = 0.1$$

In the numerical example of Sec. 9.3, c was shown to be equal to 2.64. Thus we see that in the present example, the likelihood-ratio decision rule is the same as the minimax decision rule found in Sec. 9.3.

Example 2. $X_1, \ldots, X_{16}$ are independent, each with the same normal distribution with standard deviation equal to 2 and unknown mean θ. Group I consists of the single distribution given by $\theta = 3$. $\alpha = 0.05$.

$$\max_{\theta \text{ in I}} f(x_1, \ldots, x_{16}; \theta) = f(x_1, \ldots, x_{16}; 3) = \left(\frac{1}{2\sqrt{2\pi}}\right)^{16} \exp\left[-\frac{1}{8}\sum_1^{16}(x_i - 3)^2\right]$$

Denoting $(1/16)(x_1 + \cdots + x_{16})$ by z, we know from Example 1 of Sec. 9.10 that

$$\max_\theta f(x_1, \ldots, x_{16}; \theta) = \left(\frac{1}{2\sqrt{2\pi}}\right)^{16} \exp\left[-\frac{1}{8}\sum_{i=1}^{16}(x_i - z)^2\right]$$

Therefore the likelihood ratio is equal to $\exp[-2(z - 3)^2]$, and since it decreases as $|z - 3|$ increases, it is clear that the likelihood-ratio rule in this case chooses $D = 1$ when $|z - 3| < c$, where c is a value chosen so that $r(3;s) = 0.05$. Table 1 in the Appendix shows that the proper value of c is equal to 0.98. Whether this rule is minimax depends on how group II is chosen. The rule is not minimax for the numerical example in Sec. 9.4.

Example 3. $X_1, \ldots, X_r, X_{r+1}, \ldots, X_m$ are all independent, each with a normal distribution with the same standard deviation θ_{r+1}, which is unknown. The means of $X_{r+1}, \ldots, X_m$ are known to be zero; the mean of X_i is an unknown quantity θ_i for $i = 1, \ldots, r$. Thus

$$f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_r, \theta_{r+1}) = \left(\frac{1}{\theta_{r+1}\sqrt{2\pi}}\right)^m \times \exp\left\{-\frac{1}{2\theta_{r+1}^2}\left[\sum_{i=1}^r (x_i - \theta_i)^2 + \sum_{i=r+1}^m x_i^2\right]\right\}$$

Group I consists of the distributions with $\theta_1 = \cdots = \theta_r = 0$. To maximize $f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_r, \theta_{r+1})$ with respect to $\theta_1, \ldots, \theta_r$, we set $\theta_1 = x_1, \ldots, \theta_r = x_r$. Then we have to maximize

$$\left(\frac{1}{\theta_{r+1}\sqrt{2\pi}}\right)^m \exp\left(-\frac{1}{2\theta_{r+1}^2}\sum_{i=r+1}^m x_i^2\right)$$

with respect to θ_{r+1}. As in Example 2 of Sec. 9.10, we find that the maximizing value of θ_{r+1} is $\sqrt{\frac{1}{m}\sum_{i=r+1}^m x_i^2}$, and therefore

$$\max_{\theta_1, \ldots, \theta_{r+1}} f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_r, \theta_{r+1}) = \left(\frac{2\pi}{m} e \sum_{i=r+1}^m x_i^2\right)^{-\frac{m}{2}}$$

To find $\max\limits_{\theta \text{ in } I} f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_r, \theta_{r+1})$, we note that $\theta_1 = \cdots = \theta_r = 0$, and so we must maximize

$$\left(\frac{1}{\theta_{r+1}\sqrt{2\pi}}\right)^m \exp\left(-\frac{1}{2\theta_{r+1}^2}\sum_{i=1}^m x_i^2\right)$$

with respect to θ_{r+1}. The maximizing value of θ_{r+1} is $\sqrt{\frac{1}{m}\sum\limits_{i=1}^m x_i^2}$, and therefore

$$\max_{\theta \text{ in } I} f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_{r+1}) = \left(\frac{2\pi}{m} e \sum_{i=1}^m x_i^2\right)^{-\frac{m}{2}}$$

The likelihood ratio is then equal to

$$\left(\frac{x_1^2 + \cdots + x_m^2}{x_{r+1}^2 + \cdots + x_m^2}\right)^{-\frac{m}{2}}$$

Defining T as

$$\frac{m-r}{r} \frac{X_1^2 + \cdots + X_r^2}{X_{r+1}^2 + \cdots + X_m^2}$$

it can be checked that T is a strictly decreasing function of the likelihood ratio. Therefore the likelihood-ratio decision rule is equivalent to choosing $D = 1$ when $T > c$, where c is a constant chosen so that

$$\max_{\theta \text{ in } I} r(\theta; s) = \alpha$$

Since T can be written as

$$\frac{m-r}{r} \frac{(X_1/\theta_{r+1})^2 + \cdots + (X_r/\theta_{r+1})^2}{(X_{r+1}/\theta_{r+1})^2 + \cdots + (X_m/\theta_{r+1})^2}$$

and since the standard deviation of X_i/θ_{r+1} is equal to 1, we know from Sec. 4.11 that when $\theta_1 = \cdots = \theta_r = 0$, T has an F distribution with r degrees of freedom in the numerator and $m - r$ degrees of freedom in the denominator. This fact enables us to find the value of c from Table 4 in the Appendix. When $\theta_1, \ldots, \theta_r$ are not all zero, we know from Sec. 4.12 that T has a noncentral F distribution with r degrees of freedom in the numerator, $m - r$ degrees of freedom in the denominator, and noncentrality parameter $(\theta_1^2 + \cdots + \theta_r^2)/\theta_{r+1}^2$.

The problem discussed in Example 3 is the simplest representative of an important class of problems known as "analysis-of-variance" problems. Most texts on conventional statistical methods contain detailed descriptions of such problems.

How good is the likelihood-ratio decision rule? The situation is much the same as in the case of maximum-likelihood estimates. In some problems the likelihood-ratio decision rule is a good decision rule; in other problems it is inadmissible.

EXERCISES

Section 1.3

1. If C, D, E are events, the event (C or D or E) is defined as the event which occurs on any trial where at least one of the events C, D, E occurs. The event (C and D and E) is defined as the event which occurs on any trial where all the events C, D, and E occur. Show that for any events C, D, E,

$$\begin{aligned} P(C \text{ or } D \text{ or } E) = P(C) + P(D) + P(E) - P(C \text{ and } D) \\ - P(C \text{ and } E) - P(D \text{ and } E) + P(C \text{ and } D \text{ and } E) \end{aligned}$$

2. Prove that $P(D \text{ and } E) \leq P(D)$. Under what circumstances is $P(D \text{ and } E) = P(D)$?

3. If the top card is drawn from a well-shuffled deck, what is the probability that it is an ace or a spade?

4. If a well-balanced die is rolled, what is the probability that the face that will come up is even or greater than 3?

Section 1.4

1. If an experiment consists in thoroughly shuffling a deck of cards and turning up the top card, what is the conditional probability that the top card is a spade, given that it is a black card?

2. If a well-balanced die is rolled, what is the conditional probability that the face coming up is even, given that it is greater than 3?

3. If the top card of a well-shuffled deck is turned up, what is the conditional probability that it is an ace, given that it is a spade?

Section 1.5

1. When an experiment consists in thoroughly shuffling a deck of cards and turning up the top card, which of the following pairs of events are independent:

(*a*) Top card spade, top card black.
(*b*) Top card spade, top card red.
(*c*) Top card ace, top card black.
(*d*) Top card ace, top card above a 10.

2. If two well-balanced coins are tossed, what is the probability both will come up head? What is the probability that one will come up head and the other tail?

3. If C is independent of E and D is independent of E, is the event (C and D) necessarily independent of E? Is the event (C or D) necessarily independent of E?

4. If C is independent of D, is the event (not C) necessarily independent of D?

Section 1.6

1. If n well-balanced coins are tossed, what is the probability that they will all come up head?

2. If n well-balanced coins are tossed, what is the probability that they will all show the same face?

3. For any events $A_1, A_2, \ldots, A_n$, show that $P(A_1$ and A_2 and $\cdots$ and $A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1$ and $A_2) \cdots P(A_n \mid A_1$ and A_2 and $\cdots$ and $A_{n-1})$.

4. Find an example where there are events D, E, F, with D, E independent, D, F independent, E, F independent, but D, E, F not mutually independent.

5. If $A_1, \ldots, A_n$, B are any events, show that the event $[(A_1$ or $\cdots$ or $A_n)$ and $B]$ is exactly the same event as $[(A_1$ and $B)$ or $\cdots$ or $(A_n$ and $B)]$.

6. If $A_1, \ldots, A_n$ are mutually exclusive by pairs and B is any event, show that the events $(A_1$ and $B), \ldots, (A_n$ and $B)$ are mutually exclusive by pairs.

7. If $A_1, \ldots, A_n$ are mutually exclusive by pairs, show that $P(A_1$ or A_2 or $\cdots$ or $A_n \mid B) = P(A_1 \mid B) + P(A_2 \mid B) + \cdots + P(A_n \mid B)$.

8. Show that $P(A$ or $B \mid C) = P(A \mid C) + P(B \mid C) - P(A$ and $B \mid C)$.

Section 1.8

1. Write out a formal proof of the fact that the number of different ways of choosing k places out of m places is $m!/[k!(m - k)!]$.

2. Prove that the number of different ways of dividing m places into r different groups, with exactly k_i places in group i, where $k_1 + k_2 + \cdots + k_r = m$, is $m!/(k_1!\,k_2! \cdots k_r!)$.

3. Suppose an experiment has r possible outcomes, with the probability of the outcome of type i equal to p_i, where $p_1 + \cdots + p_r = 1$. If the experiment is performed m separate times, prove that

P (outcome of type 1 occurs exactly k_1 times, and
outcome of type 2 occurs exactly k_2 times, and $\cdots$
and outcome of type r occurs exactly k_r times)

$$= \frac{m!}{k_1!\,k_2! \cdots k_r!}\, p_1^{k_1}\, p_2^{k_2} \cdots p_r^{k_r}$$

where $k_1, k_2, \ldots, k_r$ are any nonnegative integers with $k_1 + k_2 + \cdots + k_r = m$. (Use the formula of Exercise 2 and the reasoning of Example 1.)

4. Find the probability that the top five cards of a well-shuffled deck will be of the same unspecified suit.

5. Find the probability that the top five cards of a well-shuffled deck contain four cards of the same unspecified denomination.

6. Find the probability that the top five cards of a well-shuffled deck contain cards of five different denominations.

7. If Example 3 is generalized in the sense that there are r_{I} red cards in box I, b_{I} black cards in box I, r_{II} red cards in box II, b_{II} black cards in box II, find the formula for the probability that the card drawn from box II is red.

8. If Exercise 7 is modified by having two cards transferred from box I, find the formula for the probability that the card drawn from box II is red.

9. If cards are drawn one by one from a well-shuffled deck, what is the probability that the sixth card drawn is the second heart drawn?

10. If a fair die is rolled twice, what is the probability that the total number of spots will be 7?

11. If a fair die is rolled eight times, what is the probability that three times a 1-spot will appear, three times a 2-spot will appear, and twice a 5-spot will appear?

12. If a fair die is rolled until a 1-spot appears, find the probability that fewer than 5 rolls will be required.

Section 2.2

1. If two fair dice are thrown, what is the probability distribution for the chance variable defined as the total number of spots that will be face up?

2. If a coin has probability $1/3$ of coming up head on each toss, and it is tossed three times, what is the probability distribution of the chance variable defined as the number of throws on which a head will come up?

3. If three fair dice are rolled, what is the probability distribution for the chance variable defined as the largest of the three numbers that will be face up?

4. If a hand of three cards is dealt from a well-shuffled deck, what is the probability distribution for the chance variable defined as the number of spades in the hand?

Section 2.3

1. If $r(x)$ and $s(x)$ are any functions of x, and A and B are any constants, show that $E\{Ar(X) + Bs(X)\} = A\,E\{r(X)\} + B\,E\{s(X)\}$.

2. If c is a constant, prove that $E\{(X - c)^2\} = E\{X^2\} - 2cE\{X\} + c^2$.

3. What constant value of c makes $E\{(X - c)^2\}$ a minimum?

4. If X denotes the chance variable defined in Exercise 1 of Sec. 2.2, compute $E\{X\}$ and $E\{X^2\}$.

5. Suppose a magazine dealer buys a certain magazine from the publisher at 20 cents per copy, sells to customers at 50 cents per copy, and returns unsold copies to the publisher for 10 cents per copy. There are 10 potential customers for the magazine, and each has probability $1/3$ of actually requesting the magazine. If the dealer stocks 5 copies of the magazine, what is the expected value of his net profit? What number of copies should he stock to maximize his expected net profit?

6. If X denotes the chance variable defined in Exercise 3 of Sec. 2.2, find $E\{1/X\}$, $E\{2^X\}$.

7. If X denotes the chance variable defined in Exercise 4 of Sec. 2.2, find $E\{X\}$, $E\{(X - E\{X\})^2\}$.

8. Find a chance variable X with a probability distribution such that $E\{X^2\} = (E\{X\})^2$.

Section 2.4

1. Draw the graph of the cumulative distribution function for the chance variable of Exercise 1 of Sec. 2.2. Find $P(2 < X < 6)$ from the graph.

2. Draw the graph of the cumulative distribution function for the chance variable of Exercise 2 of Sec. 2.2.

3. Draw the graph of the cumulative distribution function for the chance variable of Exercise 3 of Sec. 2.2. From the graph, find $P(3 < X \leq 6)$.

4. Draw the graph of the cumulative distribution function of Exercise 4 of Sec. 2.2.

Section 2.6

1. Suppose a fair coin is tossed 5 times. Let X denote the number of times a head will come up, and Y denote the number of times a tail will come up. Construct the table of the joint probability distribution for X, Y.

2. Suppose we construct a coin with probability $1/3$ of coming up head, $1/3$ of coming up tail, and probability $1/3$ of standing on edge. The coin is tossed 5 times. Let X denote the number of times a head will come up, and let Y denote the number

of times a tail will come up. Construct the table of the joint probability distribution for X, Y.

3. Suppose four fair dice are rolled. Let X denote the smallest number that will come up, and let Y denote the largest. Construct the table of the joint probability distribution for X, Y.

4. Suppose a hand of four cards is dealt from a well-shuffled deck. Let X denote the number of spades in the hand and Y denote the number of hearts in the hand. Construct the table of the joint probability distribution for X, Y.

Section 2.7

1. For the chance variables X, Y defined in Exercise 1 of Sec. 2.6, compute $E\{X\}$, $E\{Y\}$, $E\{2X + 3Y\}$, $E\{XY\}$, $E\{X^2\}$, $E\{(X + Y)^2\}$, $E\{X^2 + Y^2\}$.

2. In Exercise 2 of Sec. 2.6, define Z as the number of times the coin will stand on edge. Compute $E\{XYZ\}$, $E\{X + Y + Z\}$, $E\{X^2YZ^2\}$.

3. For the chance variables X, Y defined in Exercise 3 of Sec. 2.6, compute $E\{X - Y\}$, $E\{XY\}$, $E\{1/XY\}$.

4. For the chance variables X, Y defined in Exercise 4 of Sec. 2.6, compute $E\{X^Y\}$, $E\{3^{XY}\}$, $E\{X2^Y\}$.

Section 2.8

1. For the chance variables X, Y defined in Exercise 1 of Sec. 2.6, compute $F(x,y)$ for the 36 pairs of values x,y ranging over the integers from 0 to 5.

2. For the chance variables X, Y defined in Exercise 2 of Sec. 2.6, compute $P(0 < X < 2, 0 < Y < 2)$.

3. For the chance variables defined in Exercise 3 of Sec. 2.6, compute $F(x,y)$ for all combinations x, y appearing in the headings of the tabled probability distribution.

Section 2.9

1. For the chance variables X, Y defined in Exercise 1 of Sec. 2.6, find the marginal distributions for X and Y in table form, and plot the marginal cdf's for X and Y.

2. For the chance variables X, Y, Z defined in Exercise 2 of Sec. 2.7, find the joint marginal distribution for X, Y in table form.

3. For the chance variables defined in Exercise 3 of Sec. 2.6, find the marginal distribution for X and the marginal distribution for Y, in table form.

4. Prove that $E\{r(X)\}$ is the same whether it is computed from the joint probability distribution for X, Y or from the marginal distribution for X, where $r(x)$ is any function.

Section 2.10

1. For the chance variables X, Y defined in Exercise 1 of Sec. 2.6, find the conditional distribution for X given that $Y = 2$.

2. For the chance variables X, Y, Z defined in Exercise 2 of Sec. 2.7, find the joint conditional distribution (in table form) for X, Y, given that $Z = 1$. Find the conditional distribution for X, given that $Y + Z = 2$.

3. Suppose that X, Y are any jointly distributed chance variables. Suppose that each of the symbols $S_1, \ldots, S_k$ denotes a certain set of possible values for Y, and A_i denotes the event (Y takes a value in S_i). Show that if the events $A_1, \ldots, A_k$ are mutually exclusive by pairs and $P(A_1) + P(A_2) + \cdots + P(A_k) = 1$, then $E\{X\} = P(A_1)E\{X \mid A_1\} + P(A_2)E\{X \mid A_2\} + \cdots + P(A_k)E\{X \mid A_k\}$.

4. For the chance variables X, Y defined in Exercise 3 of Sec. 2.6, find the conditional distribution for Y, given that $X = 2$. Find $E\{Y \mid X = 2\}$.

Section 2.11

1. If X, Y are independent, show that $E\{(X - E\{X\})(Y - E\{Y\})\} = 0$.

2. Construct a joint probability distribution for chance variables X, Y such that $E\{XY\} = E\{X\}E\{Y\}$ but X, Y are not independent.

3. Prove directly that if $F(x,y) = F_1(x)F_2(y)$ for all values of x, y, then $P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$ for all values of x, y.

4. If the set $X_1, \ldots, X_r$ is independent of the set $Y_1, \ldots, Y_s$, prove that $E\{g(X_1, \ldots, X_r)h(Y_1, \ldots, Y_s)\} = E\{g(X_1, \ldots, X_r)\}E\{h(Y_1, \ldots, Y_s)\}$, for any functions $g(x_1, \ldots, x_r)$, $h(y_1, \ldots, y_s)$.

Section 3.1

1. If a fair die is rolled until a 6-spot appears, and X is defined as the number of rolls that will be required, what is the probability distribution of X?

2. If a fair die is rolled until a 6-spot appears on two successive rolls, and X is defined as the number of rolls that will be necessary, what is the probability distribution of X?

Section 3.3

1. Describe the scale S which would give as the cdf for X the function $F(x)$ defined as follows:

$$\begin{array}{ll} F(x) = 0 & \text{for } x \leq -1 \\ F(x) = (1/4)(x + 1)^2 & \text{for } -1 < x < 1 \\ F(x) = 1 & \text{for } x \geq 1 \end{array}$$

2. Describe the scale S which would give as the cdf for X the function $F(x)$ defined as follows:

$$\begin{array}{ll} F(x) = \dfrac{1}{1 + x^2} & \text{for } x < 0 \\[2ex] F(x) = 1 & \text{for } x \geq 0 \end{array}$$

Section 3.4

1. Find the pdf corresponding to the cdf of Exercise 1 of Sec. 3.3.

2. Find the pdf corresponding to the cdf of Exercise 2 of Sec. 3.3.

3. Find the cdf corresponding to the pdf $f(x)$ defined as follows:

$$\begin{array}{ll} f(x) = 0 & \text{for } x < 0 \\ f(x) = x & \text{for } 0 < x < 1 \\ f(x) = 2 - x & \text{for } 1 < x < 2 \\ f(x) = 0 & \text{for } x > 2 \end{array}$$

4. Which of the following functions $f(x)$ are probability density functions:

(*a*) $f(x) = 0$ for $x < 0$, $f(x) = e^{-x}$ for $x > 0$

(*b*) $f(x) = (1/2)e^{-|x|}$ for $-\infty < x < \infty$

(*c*) $f(x) = \dfrac{1}{1 + |x|}$ for $-\infty < x < \infty$

(*d*) $f(x) = 2e^{-2x}$ for $-\infty < x < \infty$

(*e*) $f(x) = \dfrac{1}{\pi}\dfrac{1}{1 + x^2}$ for $-\infty < x < \infty$

Section 3.5

1. Suppose a well-balanced arrowhead spinner is mounted on a round dial labeled from 0 to 1. Let X be the number to which the arrowhead will point. (If

the arrowhead points to the place which could be either 0 or 1, it is to be read as 1.) If only k decimal places can be read on the dial, compute $E\{X\}$, $E\{X^2\}$. What do these approach as k increases?

2. For the pdf in Exercise 3 of Sec. 3.4, compute $E\{X^r\}$.

3. For each $f(x)$ in Exercise 4 of Sec. 3.4 which is a pdf, compute $E\{X\}$.

Section 3.6

1. Suppose we have two well-balanced arrowheaded spinners, each mounted as in Exercise 1 of Sec. 3.5. One of the dials can be read to two decimal places, the other can be read to an infinite number of decimal places. One of the spinners is chosen at random and set in motion, and the chance variable X is defined as the number to which the spinner will point. What is the cdf for X? Compute $E\{X\}$, $E\{X^2\}$.

Section 3.7

1. Suppose that the number of copies of a magazine that will be requested from a newsstand is a chance variable X, where

$$P(X = x) = \frac{10!}{x!\,(10-x)!}(\tfrac{3}{4})^x(\tfrac{1}{4})^{10-x} \qquad \text{for } x = 0, 1, \ldots, 10$$

If the dealer makes a net profit of 50 cents on each magazine sold and a net "profit" of -20 cents on each magazine unsold, what is the probability distribution of his net profit if he stocks 6 magazines? What number should he stock to maximize the expected value of his net profit?

2. Suppose that a tax on gross income described as follows is imposed:

Gross income	*Percentage of income taken*
Above 0, but less than 5,000	10
5,000 or above, less than 10,000	15
10,000 or above, less than 25,000	20
25,000 or above	40

If gross income is a chance variable X with pdf $f(x) = 0$ for $x < 0$, $f(x) = 0.0001e^{-0.0001x}$ for $x > 0$, find the pdf for income after tax is taken out.

3. If X has cdf $1 - e^{-x}$ for $x > 0$, find the cdf for the chance variable Y defined as $1 - e^{-X}$.

4. If X has cdf x^k for $0 \leq x \leq 1$, where k is a given positive number, find the cdf for the chance variable Y defined as X^k.

5. If X has pdf $f(x) = 1$ for $0 < x < 1$, $f(x) = 0$ for $x < 0$ or $x > 1$, find the pdf for the chance variable Y defined as $-\log X$.

Section 3.8

1. Suppose X, Y have the joint pdf $f(x,y) = 1$ for $0 < x < 1$ and $0 < y < 1$, $f(x,y) = 0$ for $x < 0$ or $x > 1$ or $y < 0$ or $y > 1$. Find the cdf $F(x,y)$. Find $P(0 < X \leq \tfrac{1}{4}, \tfrac{1}{4} < Y \leq \tfrac{1}{2})$, $P(\tfrac{1}{2} < X + Y < 1)$, $E\{X\}$, $E\{X + Y\}$, $E\{X^2 + Y^2\}$, $E\{XY\}$.

2. Suppose X, Y have the joint pdf $f(x,y) = 1/\pi$ for $x^2 + y^2 < 1$, $f(x,y) = 0$ for $x^2 + y^2 > 1$. Find the cdf $F(x,y)$. Find $P(-\tfrac{1}{4} < X \leq 0, 0 \leq Y \leq 1)$, $P(X + Y < -\tfrac{1}{4})$, $E\{X\}$, $E\{X + Y\}$, $E\{X^2 + Y^2\}$, $E\{XY\}$.

3. Suppose that X, Y have the joint pdf $f(x,y) = 2$ for $x > 0$, $y > 0$, and $x + y < 1$, $f(x,y) = 0$ for $x < 0$ or $y < 0$ or $x + y > 1$. Find the cdf $F(x,y)$. Find $P(\tfrac{1}{2} < X \leq \tfrac{1}{3}, 0 < Y \leq \tfrac{1}{4})$, $E\{X\}$, $E\{Y\}$, $E\{XY\}$.

4. Suppose X, Y have the joint pdf $f(x,y) = 3(x + y)$ for $x > 0$, $y > 0$, and $x + y < 1$, $f(x,y) = 0$ for $x < 0$ or $y < 0$ or $x + y > 1$. Find the cdf $F(x,y)$. Find $P(0 < X < \frac{1}{4}, \frac{1}{4} < Y < \frac{1}{2})$, $E\{X - Y\}$, $E\{(X + Y)^2\}$, $E\{X^2 + Y^2\}$.

Section 3.9

1. If X, Y have the joint pdf described in Exercise 1 of Sec. 3.8, find the joint pdf for $W = X + Y$, $Z = X - Y$.

2. If X, Y have the joint pdf described in Exercise 2 of Sec. 3.8, find the joint pdf for $W = X + Y$, $Z = X$.

3. If X, Y have the joint pdf described in Exercise 3 of Sec. 3.8, find the joint pdf for $W = X + Y$, $Z = X - Y$.

4. If X, Y have the joint pdf described in Exercise 4 of Sec. 3.8, find the joint pdf for $W = X + Y$, $Z = X + 2Y$.

Section 3.10

1. If X, Y have the joint pdf described in Exercise 1 of Sec. 3.8, find the marginal cdf and pdf for X and the marginal cdf and pdf for Y. Are X, Y independent chance variables? Find the conditional pdf for X given that $Y = \frac{1}{2}$, and $E\{X \mid Y = \frac{1}{2}\}$.

2. If X, Y have the joint pdf described in Exercise 2 of Sec. 3.8, find the marginal cdf and pdf for X and the marginal cdf and pdf for Y. Are X, Y independent chance variables? Find the conditional pdf for X given that $Y = 0$. Find the conditional pdf for X given that $Y < 0$. Find $E\{X^2 \mid Y = 0\}$, $E\{X \mid Y \leq 0\}$.

3. If X, Y have the joint pdf described in Exercise 3 of Sec. 3.8, find the marginal cdf and pdf for X and the marginal cdf and pdf for Y. Are X, Y independent chance variables? Find the conditional cdf for X given that $Y = \frac{1}{2}$.

4. If X, Y have the joint pdf described in Exercise 4 of Sec. 3.8, find the marginal pdf and cdf for X and the marginal pdf and cdf for Y. Are X, Y independent chance variables? Find the conditional pdf for Y given that $X = \frac{1}{2}$. Find $E\{Y \mid X = \frac{1}{2}\}$.

Section 3.11

1. Develop a formula for $P(x_1 < X \leq x_2, y_1 < Y \leq y_2, z_1 < Z \leq z_2)$ in terms of the eight quantities $F(x_1,y_1,z_1)$, $F(x_1,y_1,z_2)$, . . . , $F(x_2,y_2,z_2)$.

2. If X, Y, Z have joint pdf $f(x,y,z) = \frac{1}{6}$ for $x > 0$, $y > 0$, $z > 0$, and $x + y + z < 1$, and $f(x,y,z) = 0$ for $x < 0$ or $y < 0$ or $z < 0$ or $x + y + z > 1$, find the pdf for $W = X + Y + Z$ by two different methods:

(*a*) The method described in the text of introducing convenient "extra" chance variables.

(*b*) Find $\displaystyle\iiint\limits_{x+y+z \leq w} f(x,y,z)\, dx\, dy\, dz$, and differentiate with respect to w.

3. For the chance variables X, Y, Z described in Exercise 2, find $f_{1,2}(x,y)$, $f_2(y)$, $E\{XYZ\}$, $E\{XY \mid Z = \frac{1}{2}\}$.

4. If the joint pdf for X, Y, Z is $f(x,y,z) = e^{-(x+y+z)}$ for $x > 0$, $y > 0$, $z > 0$, $f(x,y,z) = 0$ for $x < 0$ or $y < 0$ or $z < 0$, find the pdf for $W = X + Y + Z$ by first finding $\displaystyle\iiint\limits_{x+y+z \leq w} f(x,y,z)\, dx\, dy\, dz$.

5. If $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are jointly distributed chance variables, where the set of chance variables $X_1, \ldots, X_m$ is independent of the set of chance variables

$Y_1, \ldots, Y_n$, and if the chance variable W is defined as a function of $X_1, \ldots, X_m$ while the chance variable Z is defined as a function of $Y_1, \ldots, Y_n$, prove that W, Z are independent.

6. For each of the joint distributions described in the exercises of Sec. 3.8, find the pdf for $W = X + Y$.

Section 4.1

1. If X has pdf $f(x) = 1$ for $0 < x < 1$, $f(x) = 0$ for $x < 0$ or $x > 1$, find $E\{X\}$, $E\{X^2\}$, $E\{X^3\}$ by direct integration and also by finding $M_X(t)$ and differentiating.

2. If X has the probability distribution given by the following table:

Possible values	1	2	3	4	5
Probability	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

find $E\{X\}$, $E\{X^2\}$, $E\{X^3\}$ by direct computation and also by finding $M_X(t)$ and differentiating.

3. If X has pdf $f(x) = (1/\pi)(1 + x^2)^{-1}$, which moments of X exist? For which values of t does $M_X(t)$ exist?

Section 4.2

1. For each of the following probability distributions, state whether or not it is a binomial distribution. If it is, state what the parameters are.

(*a*)

0	1	2	3
$\frac{1}{27}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{8}{27}$

(*b*)

0	1	2
$\frac{1}{16}$	$\frac{7}{16}$	$\frac{8}{16}$

2. If X has a binomial distribution with parameters 4, 0.4, find $P(X = 2)$.

3. Show that if $X_1, X_2, \ldots, X_r$ are independent chance variables, each with a binomial distribution with the parameters for X_i given by n_i, p, then $Z = X_1 + \cdots + X_r$ has a binomial distribution. What are the parameters for Z?

4. If X has a binomial distribution with parameters n, p, find $E\{X^3\}$, $E\{X^4\}$.

Section 4.3

1. If on the average 1 person in 10,000 contracts a certain noncontagious disease in one year, what is the probability that more than 3 people in a town of 20,000 contract the disease in a given year?

2. If X has a binomial distribution with $n = 100$ and $p = 0.01$, find $P(X = 0)$, $P(X = 1)$, $P(X = 2)$. Compare these values with the corresponding probabilities given by the Poisson distribution with parameter 1.

3. If X has a Poisson distribution with parameter λ, find $E\{X^3\}$, $E\{X^4\}$.

4. If $X_1, \ldots, X_r$ are independent chance variables, each with a Poisson distribution, with the parameter for X_i equal to λ_i, show that $Z = X_1 + \cdots + X_r$ has a Poisson distribution. What is the parameter for Z?

5. In each of the following cases, describe why the chance variable could reasonably be supposed to have a Poisson distribution:

(*a*) X is the number of typographical errors on a given page of a book, when the page contains a large number of symbols.

(*b*) X is the number of telephone calls originated during a given brief time period in a large city.

(*c*) X is the number of defects in a given piece of cloth.

6. Suppose the probability that the ith person contracts a noncontagious disease in a given time period is p_i. Show that if there are n people, and if $p_i \leq K/n$ for all i, and $p_1 + \cdots + p_n = \lambda$, then as n increases, the distribution of the total number of people who will contract the disease in the given period approaches a Poisson distribution with parameter λ.

Section 4.4

1. If X has a hypergeometric distribution with parameters n, R, B, find $E\{X\}$.

2. If X has a hypergeometric distribution with parameters n, R, B, show by direct calculation that

$$\lim_{R+B\to\infty} P(X = x) = \frac{n!}{x!\,(n-x)!}\left(\frac{R}{R+B}\right)^{x}\left(\frac{B}{R+B}\right)^{n-x}$$

3. If a batch of 100 items contains 5 defectives and a sample of 6 items is drawn from the batch, compute the probability distribution for the number of defectives in the sample.

4. If X has a hypergeometric distribution with parameters n, R, B, what conditions on the parameters would imply that the distribution of X is approximately Poisson?

Section 4.5

1. If X has a uniform distribution between A and B, find $M_X(t)$.

2. If X has a uniform distribution between A and B, and C, D are constants, what is the distribution of $Y = (X - C)/D$?

3. Construct a joint distribution for X, Y so that X, Y are not independent and each of the marginal distributions is a uniform distribution between 0 and 1.

4. Suppose X has cdf $F(x)$, where $F(x)$ is not continuous. What can be said about the cdf for $Y = F(X)$?

Section 4.6

1. Suppose the number of items that will be demanded has a normal distribution with mean 10,000 and standard deviation 1,000. What is the probability that more than 11,000 items will be demanded? That fewer than 8,000 items will be demanded? That between 9,000 and 12,000 items will be demanded?

2. Suppose the demand is as described in Exercise 1, and a net profit of \$1 is made on each item sold, a net loss of 50 cents is made on each item unsold. How many items should be stocked to maximize expected value of net profit?

3. Suppose that the situation in Exercise 2 is modified by the imposition of a graduated tax on net profit, as follows. If net profit is equal to n, the percentage of tax is $100(1 - e^{-n})$. How many items should be stocked to maximize the expected value of net profit after taxation?

4. If X has a normal distribution with parameters μ, σ, find $E\{X^3\}$, $E\{X^4\}$.

5. Evaluate $(1/\sqrt{2\pi})\int_{-\infty}^{1} \exp(-(\frac{1}{2})y^2)\,dy$ approximately by using the fact that it is equal to $\frac{1}{2} + (1/\sqrt{2\pi})\int_{0}^{1} \exp(-(\frac{1}{2})y^2)\,dy$ and approximating $\exp(-(\frac{1}{2})y^2)$ in this last expression by $1 - (\frac{1}{2})y^2 + y^4/8 - y^6/48 + y^8/384$. Compare this approximate value with the value in Table 1 in the Appendix.

Section 4.7

1. If X has a Poisson distribution with parameter λ, prove that the distribution of $(X - \lambda)/\sqrt{\lambda}$ approaches the standard normal distribution as λ increases.

2. If X, Y are independent, each with a normal distribution with parameters μ_1, σ_1 and μ_2, σ_2, respectively, find the distribution of $X + Y$.

3. If X has a binomial distribution with parameters n, p, prove that the moment generating function for $(X - np)/\sqrt{np(1 - p)}$ approaches the moment generating function for a standard normal distribution as n increases.

4. Carry out the proof of the central limit theorem without assuming that $\sigma_1^2 = \sigma_2^2 = \cdots$, and $\Delta_1(t) = \Delta_2(t) = \cdots$.

5. If $X_1, X_2, \ldots, X_n$ are independent chance variables, each with a uniform distribution between 0 and 1, define Y_n as $(X_1 + \cdots + X_n - (\frac{1}{2})n)/\sqrt{n/12}$, and find the moment generating function of Y_n. What does this moment generating function approach as n increases?

Section 4.8

1. If W has a chi-square distribution with n degrees of freedom, show that the moment generating function for $(W - n)/\sqrt{2n}$ approaches the moment generating function for the standard normal distribution, as n increases.

2. If W has a chi-square distribution with 2 degrees of freedom, find the cdf for W.

3. Suppose that W has a chi-square distribution with 30 degrees of freedom. Using the fact that $(W - 30)/\sqrt{60}$ has approximately a standard normal distribution, approximate the value of w for which $P(W \leq w) = A$, for the values of A given in Table 2 in the Appendix. Compare these approximate values with the exact values from Table 2 in the Appendix.

Section 4.9

1. Give a complete proof that if W has a noncentral chi-square distribution with parameters n and m, then $P(W \leq w)$ decreases as m increases.

2. If W has a noncentral chi-square distribution with parameters n and m, find $E\{W\}$, $E\{W^2\}$, $E\{W^3\}$.

3. Suppose W has a noncentral chi-square distribution with parameters n and m. Denote the mean of W by A and the standard deviation of W by B. Show that as n increases, the moment generating function for $(W - A)/B$ approaches the moment generating function for the standard normal distribution.

Section 4.10

1. If T has a t distribution with n degrees of freedom, show that the pdf for T approaches the standard normal pdf as n increases.

2. If T has a t distribution with n degrees of freedom, find $E\{T\}$, $E\{T^2\}$.

3. If T has a t distribution with 1 degree of freedom, find and plot the cdf for T.

4. If T has a t distribution with 3 degrees of freedom, find and plot the cdf for T.

Section 4.11

1. If T has a t distribution with n degrees of freedom, show that T^2 has an F distribution. What are the parameters of this F distribution?

2. If Z has an F distribution with r degrees of freedom in the numerator and s degrees of freedom in the denominator, show that $1/Z$ has an F distribution. What are the parameters of this F distribution?

3. If Z has an F distribution with r degrees of freedom in the numerator and s degrees of freedom in the denominator, find the mean and variance of Z.

4. If Z has an F distribution with r degrees of freedom in the numerator and s degrees of freedom in the denominator, which moments of Z exist?

Section 4.12

1. If Z has a noncentral F distribution with noncentrality parameter m, show that $P(Z < z)$ decreases as m increases.

Section 4.13

1. Find the correlation coefficient between X and Y for each of the following joint pdf's:

(a) $f(x,y) = 2$ for $x > 0$, $y > 0$, and $x + y < 1$
$f(x,y) = 0$ for $x < 0$, or $y < 0$, or $x + y > 1$
(b) $f(x,y) = 1/\pi$ for $x^2 + y^2 < 1$
$f(x,y) = 0$ for $x^2 + y^2 > 1$
(c) $f(x,y) = e^{-(x+y)}$ for $x > 0$ and $y > 0$
$f(x,y) = 0$ for $x < 0$ or $y < 0$

2. Construct a joint probability distribution in table form for chance variables X, Y, such that $\rho_{XY} = 1$.

3. Find $E\{X^2Y^2\}$ for each of the joint pdf's in Exercise 1.

4. Find the correlation coefficient between X and Y for the joint pdf described in Exercise 4 of Sec. 3.8.

Section 4.14

1. For each of the joint distributions in Exercise 1 of Sec. 4.13, find the joint moment generating function, $E\{XY\}$, $E\{X\}$, $E\{X^2\}$, $E\{Y\}$, $E\{Y^2\}$.

2. From the joint moment generating function for $X_1, \ldots, X_m$, how is the moment generating function for $X_1 + \cdots + X_m$ obtained? Illustrate with an example.

3. Find the joint moment generating function for the joint distribution described in Exercise 4 of Sec. 3.8.

Section 4.16

1. If X_1, X_2 have a bivariate normal distribution with parameters $u_1, u_2, \sigma_1, \sigma_2, \rho$, find the conditional pdf for X_1 given that $X_2 = t$.

2. If X_1, X_2 have a bivariate normal distribution with parameters $u_1, u_2, \sigma_1, \sigma_2, \rho$, find the joint pdf for $Y_1 = A_1X_1 + B_1X_2 + C_1$, $Y_2 = A_2X_1 + B_2X_2 + C_2$, where $A_1, A_2, B_1, B_2, C_1, C_2$ are constants with $A_1B_2 \neq A_2B_1$.

3. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are n pairs of chance variables, each pair being independent of all the other pairs and the joint distribution of (X_i, Y_i) being the same for all values of i. Denote the mean of X_i by A, the mean of Y_i by B, the standard deviation of X_i by C, and the standard deviation of Y_i by D. Define W_n as $(X_1 + \cdots + X_n - nA)/\sqrt{n}C$, Z_n as $(Y_1 + \cdots + Y_n - nB)/\sqrt{n}D$. Find the joint moment generating function for (W_n,Z_n) in terms of the joint moment generating function for (X_1, Y_1). Show that under conditions analogous to those imposed in Sec. 4.7, the joint moment generating function for (W_n,Z_n) approaches a joint moment generating function for a bivariate normal distribution as n increases.

Section 4.17

1. Suppose that the probability that a telephone conversation ends during the time interval $(t, t + \Delta t)$, given that it has not ended before time t, is $0.1t\,\Delta t + q(t,\Delta t)$,

where $[q(t,\Delta t)]/\Delta t$ approaches zero as Δt approaches zero, uniformly in t. What is the probability that the conversation lasts for more than 2 min?

2. Suppose that as soon as one telephone conversation ends, another one is started, and that the total length of each telephone conversation is a chance variable with an exponential distribution with parameter θ. Let Y denote the number of conversations that will be started between time 0 and time t. Find the probability distribution for Y.

3. If X is a chance variable representing the length of life of a piece of equipment and X has pdf $f(x)$ and cdf $F(x)$, show that the "death rate" at time t is equal to $f(t)/[1 - F(t)]$.

4. If X has an exponential distribution with parameter θ, find the conditional distribution of $X - A$, given that $X > A$, where A is a positive constant.

Section 5.1

1. Write out a proof of the supporting hyperplane theorem for the case of sets in three dimensions.

Section 5.6

1. Suppose the number of items that will be demanded is a chance variable Y with a Poisson distribution with unknown parameter θ. On each item that is sold, there is a net profit of \$10; on each item remaining unsold, there is a net loss of \$5. The decision to be made is how many items to stock. Before choosing the decision, X_1, X_2, X_3 will be observed, where these are independent chance variables, independent of Y, each with a Poisson distribution with parameter θ. If the decision rule s is to stock a number of items equal to the largest integer which is not greater than $(\frac{1}{3})(X_1 + X_2 + X_3)$, what is $r(\theta;s)$?

2. Suppose a machine may be purchased for \$50,000. Its total useful life (in months) is a chance variable Y with an exponential distribution with unknown parameter θ. While the machine is operating, it yields a net profit at the rate of \$500 per month. When the machine fails, it has a salvage value of \$2,000. There are two possible decisions: to buy the machine or not to buy. Before a decision is made, the chance variables X_1, X_2 are to be observed, where X_1, X_2 are independent, independent of Y, and each has the same distribution as Y. Suppose the decision rule s is to choose that decision which would be appropriate if θ were known to be equal to $(\frac{1}{2})(X_1 + X_2)$. Find the functions $s(D;x)$, $r(\theta;s)$.

3. Suppose that a company is supposed to deliver 5 items to a customer by a certain time. On each item actually delivered, the company makes a net profit of \$500, but incurs a penalty cost of \$1,000 for each item undelivered. There is an unknown probability θ of spoilage of each item during its production. There is no time for reruns, so that only items started in production at the beginning of the time period could be finished by the delivery date. It costs \$300 for each item started through production. D denotes the number of items started, Y denotes the number of items that will survive the production process in good condition. Spoiled and surplus items have no value. Before a decision is chosen, a chance variable X will be observed, where X is the number of items surviving a production process with the probability θ of spoilage of each item, when 6 items were started through the process. Suppose the decision rule s is to choose the decision that would be appropriate if θ were known to be equal to $1 - X/6$. Find the functions $s(D;x)$, $r(\theta;s)$.

4. Modify Exercise 3 by giving surplus finished items a value of \$100 apiece.

5. In certain problems, the joint distribution for X, Y is affected by the decision D that is chosen, so that the joint distribution must be written as $f(x,y;\theta,D)$. Does

this cause any difficulty in the development? How must the formula for $r(\theta;s)$ be modified? Are any of the exercises above of this type?

6. Suppose that a rancher raising 1,000 head of cattle has to decide whether or not to include a special vitamin supplement in their diet. Including the supplement would cost \$5,000. When the animals are sold, each is classified as either Grade A or Grade B, and the price paid for each Grade A animal is \$200, the price paid for each Grade B animal is \$180. If the vitamin supplement is not included in the diet of an animal, the probability that the animal will become Grade A is $\frac{1}{2}$. If the vitamin supplement is included in the diet of an animal, the probability that the animal will become Grade A is some unknown value θ between $\frac{1}{2}$ and 1. Before making his decision, the rancher will observe (at no cost) how many of 5 test animals fed the supplement become Grade A. Letting X denote the number of the test animals that become Grade A and Y denote the number of the 1,000 animals that will become Grade A and setting $D = 1$ if the decision made is not to include the supplement, setting $D = 2$ if the decision made is to include the supplement, what is the function $W(y;D;x)$? Find $r(\theta;s)$ for the decision rule s given by $s(1;x) = 1$ for $x \leq 4$, $s(1;5) = 0$.

7. A person has to decide whether or not to insure a piece of jewelry worth \$2,000 against theft for a period of 1 year. To insure the article costs \$20. The probability that the article will be stolen during the year is an unknown value θ. Before deciding whether or not to insure the article, the person will observe (at no cost) how many articles were stolen out of 6 separate articles, each with probability θ of being stolen. $D = 1$ denotes the decision not to insure, and $D = 2$ denotes the decision to insure. Define Y as equal to 1 if the article is stolen and as equal to 0 if the article is not stolen. Define X as the number of the 6 observed articles that are stolen. What is the function $W(y;D;x)$? Find $r(\theta;s)$ for the decision rule s given by $s(1;x) = 1$ if $x \leq 3$, $s(1;x) = \frac{1}{2}$ if $x = 4,5$, $s(1;6) = 0$.

8. In Exercise 7, suppose that if the person insures the article, he puts a notice of insurance on his door which reduces the probability that the article will be stolen to $(\frac{1}{2})\theta$. Under these circumstances, answer the same questions as in Exercise 7.

9. Modify Exercise 7 by making it possible to insure the article for any amount between 0 and \$2,000. Let D denote the amount for which the article is insured. Suppose the cost of insuring the article for D dollars is equal to $0.001D$. What is the function $W(y;D;x)$? Compute $r(\theta;s)$ for the decision rule s which sets D equal to $2{,}000(X/6)$.

Section 5.7

1. Suppose Exercise 3 of Sec. 5.6 is modified so that θ is known to be either 0.1 or 0.4 and D can be no greater than 8. Graph the convex set C.

2. Suppose Exercise 4 of Sec. 5.6 is modified so that θ is known to be either 0.2 or 0.3 and D can be no greater than 7. Graph the convex set C.

3. Suppose that Exercise 6 of Sec. 5.6 is modified so that θ is known to be either $\frac{3}{4}$ or $\frac{7}{8}$. Graph the convex set C.

4. Suppose Exercise 7 of Sec. 5.6 is modified so that θ is known to be either $\frac{1}{4}$ or $\frac{3}{4}$. Graph the convex set C.

5. Suppose Exercise 8 of Sec. 5.6 is modified so that θ is known to be either 0.1 or 0.8. Graph the convex set C.

Section 5.8

1. From the diagram of Exercise 1 of Sec. 5.7, find a point representing a Bayes decision rule relative to $b(0.1) = 0.3$, $b(0.4) = 0.7$.

2. From the diagram of Exercise 2 of Sec. 5.7, find a point representing a Bayes decision rule relative to $b(0.2) = 0.5$, $b(0.3) = 0.5$.

3. From the diagram of Exercise 3 of Sec. 5.7, find a point representing a Bayes decision rule relative to $b(\frac{3}{4}) = 0.4$, $b(\frac{7}{8}) = 0.6$.

4. From the diagram of Exercise 4 of Sec. 5.7, find a point representing a Bayes decision rule relative to $b(\frac{1}{4}) = 0.8$, $b(\frac{3}{4}) = 0.2$.

5. From the diagram of Exercise 5 of Sec. 5.7, find a point representing a Bayes decision rule relative to $b(0.1) = 1$, $b(0.8) = 0$.

Section 5.9

1. For Exercise 1 of Sec. 5.7, construct a Bayes decision rule s relative to $b(0.1) = 0.3$, $b(0.4) = 0.7$. Compute $r(0.1;s)$, $r(0.4;s)$.

2. For Exercise 2 of Sec. 5.7, construct a Bayes decision rule s relative to $b(0.2) = 0.5$, $b(0.3) = 0.5$. Compute $r(0.2;s)$, $r(0.3;s)$.

3. For Exercise 3 of Sec. 5.7, construct a Bayes decision rule s relative to $b(\frac{3}{4}) = 0.4$, $b(\frac{7}{8}) = 0.6$. Compute $r(\frac{3}{4};s)$, $r(\frac{7}{8};s)$.

4. For Exercise 4 of Sec. 5.7, construct a Bayes decision rule relative to $b(\frac{1}{4}) = 0.8$, $b(\frac{3}{4}) = 0.2$. Compute the expected losses.

5. For Exercise 5 of Sec. 5.7, construct a Bayes decision rule s relative to $b(0.1) = 1$, $b(0.8) = 0$. Compute $r(0.1;s)$, $r(0.8;s)$.

6. Modify Exercise 9 of Sec. 5.6 so that the only possible values of D are 0, 1,000, 2,000 and the only possible values of θ are $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$. Construct a Bayes decision rule s relative to $b(\frac{1}{4}) = 0.1$, $b(\frac{1}{2}) = 0.6$, $b(\frac{3}{4}) = 0.3$. Compute $r(\frac{1}{4};s)$, $r(\frac{1}{2};s)$, $r(\frac{3}{4};s)$.

Section 5.10

1. In Exercise 1 of Sec. 5.6, suppose the only possible values of D are the integers from 0 to 10. Find a Bayes decision rule s relative to the a priori distribution $B(\theta)$ with pdf $b(\theta) = e^{-\theta}$ for $\theta > 0$.

2. In Exercise 2 of Sec. 5.6, find a Bayes decision rule s relative to the a priori distribution $B(\theta)$ with pdf $b(\theta) = 0.01e^{-0.01\theta}$ for $\theta > 0$.

3. In Exercise 3 of Sec. 5.6, suppose the only possible values for D are the integers from 0 to 9. Find a Bayes decision rule s relative to the a priori distribution $B(\theta)$ with pdf $b(\theta) = 1$ for $0 < \theta < 1$.

4. In Exercise 4 of Sec. 5.6, suppose the only possible values for D are the integers from 0 to 9. Find a Bayes decision rule s relative to the a priori distribution $B(\theta)$ with pdf $b(\theta) = 2\theta$ for $0 < \theta < 1$.

5. In Exercise 6 of Sec. 5.6, construct a Bayes decision rule s relative to the a priori distribution $B(\theta) = \theta^3$ for $0 < \theta < 1$. Compute $r(\theta;s)$.

6. In Exercise 7 of Sec. 5.6, construct a Bayes decision rule s relative to the a priori distribution $B(\theta) = \theta^2$ for $0 < \theta < 1$. Compute $r(\theta;s)$.

7. In Exercise 8 of Sec. 5.6, construct a Bayes decision rule relative to the a priori distribution $B(\theta) = \theta$ for $0 < \theta < 1$.

8. Modify Exercise 9 of Sec. 5.6 so that the only possible values of D are 0, 1,000, 2,000. Construct a Bayes decision rule s relative to the a priori distribution $B(\theta) = \theta^2$ for $0 < \theta < 1$. Compute $r(\theta;s)$.

Section 5.11

1. Carry out Exercise 1 of Sec. 5.10, allowing D to be any nonnegative integer.

2. Carry out Exercise 3 of Sec. 5.10, allowing D to be any nonnegative integer.

3. Carry out Exercise 4 of Sec. 5.10, allowing D to be any nonnegative integer.

4. Suppose that Y is the quantity of a continuously divisible commodity that will

be demanded. Y has a normal distribution, with mean θ_1 and standard deviation θ_2. The problem is to decide how much to stock. D can be any nonnegative number. The net profit rate is \$5 per unit sold, -2 per unit unsold. Before a decision is chosen, X_1, X_2, X_3 will be observed, where these are independent, independent of Y, and each has the same distribution as Y. Find a Bayes decision rule relative to the a priori distribution $B(\theta_1,\theta_2)$ with

$$\text{pdf } b(\theta_1,\theta_2) = \frac{0.01}{100\sqrt{2\pi}} \exp\left[-\frac{1}{2(100)^2}(\theta_1 - 10{,}000)^2 - 0.01\theta_2\right] \qquad \text{for } \theta_2 \geq 0$$

$$b(\theta_1,\theta_2) = 0 \qquad \text{for } \theta_2 < 0$$

5. For Exercise 9 of Sec. 5.6, construct a Bayes decision rule relative to the a priori distribution $B(\theta) = \theta^2$ for $0 \leq \theta \leq 1$.

6. For Exercise 9 of Sec. 5.6, construct a Bayes decision rule s relative to the a priori distribution $B(\theta)$ which has jumps equal to $1/3$ at $\theta = 0, 1/2, 1$. Compute $r(\theta;s)$.

Section 5.12

1. For Exercise 1 of Sec. 5.7, find $g_1(2,3), g_2(-2,-3), g_1(-2,4), g_2(0,1), g_1(-3,7)$.
2. For Exercise 3 of Sec. 5.7, find $g_1(1,1), g_1(0,-1), g_2(3,3), g_2(-1,3)$.

Section 5.13

1. For Exercise 1 of Sec. 5.6, find a function of X_1, X_2, X_3 which is sufficient.
2. For Exercise 2 of Sec. 5.6, find a function of X_1, X_2 which is sufficient.
3. In Exercise 6 of Sec. 5.6, X was defined as the number of test animals that became Grade A out of 5 test animals observed. More detailed information could be given by numbering the test animals and defining X_i as equal to 1 if the ith test animal becomes Grade A and as equal to 0 otherwise. Show that knowing only X is as good as knowing $X_1, \ldots, X_5$.
4. In Exercise 7 of Sec. 5.6, X was defined as the number of observed articles that were stolen. More detailed information could be given by numbering the observed articles and defining X_i as equal to 1 if the ith observed article is stolen and as equal to zero otherwise. Show that knowing only X is as good as knowing $X_1, \ldots, X_6$.

Section 5.14

1. For Exercise 1 of Sec. 5.7, find a minimax decision rule.
2. For Exercise 2 of Sec. 5.7, find a minimax decision rule.
3. If a person feels that he knows which particular value of θ is the true one, what is the appropriate a priori distribution to use? How is a Bayes decision rule relative to this a priori distribution constructed? In particular, what is the role of $X_1, \ldots, X_m$?
4. Find a minimax decision rule for each of Exercises 3, 4, 5 of Sec. 5.7.

Section 6.1

1. Suppose there are two fuel types, with the following characteristics:

	Fuel type 1	Fuel type 2
Weight per unit volume	15	10
Energy per unit volume	30	25
Cost per unit volume	1	0.8

Suppose a mixture of at least 8 units of volume must be obtained, weighing no more than 120 units and containing at least 260 units of energy. Let x_1, x_2 denote, respectively, the number of units of volume of fuel type 1 and fuel type 2 that is used. In the x_1, x_2 plane, sketch the set G of all points (x_1,x_2) satisfying the volume, weight, and energy restrictions. On this diagram, indicate all points of G with the property U. Find a point of G which minimizes the cost of the mixture.

Section 6.2

1. Suppose there are four types of food, with the following properties:

	Food type 1	2	3	4
Amt. of vitamin *A*/unit vol.	15	4	2	0
Amt. of vitamin *B*/unit vol.	6	18	1	2
Amt. of vitamin *C*/unit vol.	5	5	8	3
Amt. of vitamin *D*/unit vol.	2	1	0	7
Cost per unit volume	15	20	2	6

The problem is to obtain a diet containing at least 40 units of vitamin *A*, at least 35 units of vitamin *B*, at least 25 units of vitamin *C*, and at least 18 units of vitamin *D*. Find the quantities of the four types of food achieving these requirements at minimum cost.

2. Prove that G is a convex set.

3. Using Exercise 2, show that if we reach a tableau with no negative d's, no lowering of the objective function is possible.

4. What changes in the simplex method must be made if we want to maximize the objective function?

5. In Exercise 1, show that the convex set G has points with arbitrarily large coordinates. Why does this fail to cause any trouble in the computation? What would happen if we tried to maximize the cost?

Section 6.3

1. The proportion of residents of a special type in a large city is known to be 0.1, 0.3, or 0.5. If the proportion is 0.1, the demand for a certain product during a given time period has a normal distribution with mean 5,000 and standard deviation 100; if the proportion is 0.3, the demand for the product has a normal distribution with mean 10,000 and standard deviation 200; if the proportion is 0.5, the demand for the product has a normal distribution with mean 15,000 and standard deviation 300. The retail price of the product is \$5 per unit. A retailer can buy from the wholesaler at \$3 per unit and return any unsold product to the wholesaler at \$0.50 per unit. However, at the start of the time period, the retailer can buy 8,000 units at \$2.50 apiece, or 16,000 units at \$2 apiece. (If the retailer does either, he can still buy additional units, whenever needed, at \$3 per unit.) Thus there are three possible decisions: (1) buy 8,000 units at the start, (2) buy 16,000 units at the start, (3) do neither. Before choosing one of these three decisions, the retailer will choose four residents at random and observe how many are of the special type. Find a minimax decision rule for this problem.

2. Modify Exercise 1 by making unsold units valueless.

3. Modify Exercise 6 of Sec. 5.6 by allowing the possible values of θ to be $\frac{1}{2}, \frac{3}{4}, 1$. Find a minimax decision rule for the resulting problem.

4. Modify Exercise 7 of Sec. 5.6 by allowing the possible values of θ to be $\frac{1}{4}, \frac{1}{3}, \frac{3}{4}$. Find a minimax decision rule for the resulting problem.

5. Modify Exercise 8 of Sec. 5.6 by allowing the possible values of θ to be $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$. Find a minimax decision rule for the resulting problem.

Section 7.2

1. Find a minimax decision rule for the example in the text when it costs \$50 to observe a consumer.

2. Find a minimax decision rule for the example in the text when it costs \$150 to observe a consumer.

Section 7.5

1. Find an approximately minimax Wald sequential rule for the case $a(G_1;d_1) = 0$, $a(G_1;d_2) = 300{,}000$, $a(G_2;d_1) = 200{,}000$, $a(G_2;d_2) = 0$, $c = 1$, $h_1 = -1.1$, $h_2 = 1.1$.

Section 7.6

1. If the illustrative example in the text is modified so that θ is known to be 0.1, what decision rule should be used? If θ is known to be 0.9, what decision rule should be used?

2. If the illustrative example in the text is modified so that θ is known to be either 0.1 or 0.6, find a minimax decision rule.

3. For the illustrative example in the text, find a Bayes decision rule relative to the a priori distribution $B(\theta) = \theta$ for $0 \leq \theta \leq 1$.

4. Suppose that a newsdealer can obtain copies of a monthly magazine at only two times: at the beginning of the month and in the middle of the month. At either time, he can obtain as many copies as he wants, at 15 cents per copy. He can return any number of copies to the distributor at two times: at the middle of the month or at the end of the month. For each copy returned at the middle of the month he receives 10 cents per copy, and at the end of the month he returns all unsold copies and receives 5 cents per copy. The retail price is 30 cents per copy. The number of magazines that will be demanded from the dealer during the first half of the month and the number that will be demanded during the second half of the month are independent chance variables, each with a Poisson distribution with the same unknown parameter θ. The dealer must decide how many magazines to order at the beginning of the month and how many to order or return at the middle of the month. Before deciding, he will observe X_1, X_2, X_3, X_4, where these are independent chance variables, each with a Poisson distribution with parameter θ. Find a Bayes decision rule relative to the a priori distribution $B(\theta) = 1 - e^{-0.01\theta}$ for $\theta \geqslant 0$.

5. If Exercise 4 is modified so that θ is known to be equal to 100, what decision rule should be used?

6. Modify the example in the text so that the production process is a three-stage process and only items surviving the first two stages can enter the third stage. It costs \$50 for each item started through the third stage of production. Find a Bayes decision rule relative to the a priori distribution $B(\theta) = \theta^4$ for $0 \leqslant \theta \leqslant 1$.

7. In the construction of the Wald sequential decision rule, it was not necessary to work backward. What elements in the problem discussed in Secs. 7.3 and 7.4 allowed us to "work forward"?

Section 8.2

1. If Z_m has a binomial distribution with parameters m, p, then the central limit theorem tells us that the cdf for $(Z_m - mp)/\sqrt{mp(1-p)}$ approaches the standard normal cdf as m increases. Use this fact to approximate $P(|Z_m/m - p| \leqslant \epsilon)$ for large m. Compare the approximation to the upper bound for this probability given by Tchebycheff's inequality.

2. Suppose X is a chance variable with mean A and variance B. Prove that for any positive constant c, $P(|X - A| \leqslant c) \geqslant 1 - B/c^2$.

Section 8.3

1. If $A_1, A_2, \ldots, A_k$ are mutually exclusive by pairs, show that $P(\text{not } A_1) + P(\text{not } A_2) + \cdots + P(\text{not } A_k) \geqslant 1$.

Section 8.4

1. Write out a complete proof for Theorem 2 for the case where $F(x)$ is not necessarily continuous.

2. Suppose $X_1, \ldots, X_m$ are independent, each with the same continuous cdf $F(x)$. Define Z as $\max_x |H(x;X_1, \ldots, X_m) - F(x)|$. Define Y_i as $F(X_i)$ for $i = 1, \ldots, m$, and define W as $\max_{0 \leqslant y \leqslant 1} |H(y;\, Y_1, \ldots, Y_m) - y|$. Show that $Z = W$.

3. Use Exercise 2 to show that if $F(x)$ is continuous, the distribution of $\max_x |H(x;\, X_1, \ldots, X_m) - F(x)|$ does not depend on $F(x)$.

Section 8.5

1. Suppose the news vendor's problem is modified by imposing a tax on all net profits above zero, where the tax on a net profit of p is equal to $p(1 - e^{-Tp})$, T a given positive constant. Find the empirical decision rule for this problem.

2. In the news vendor's problem, suppose $X_1, \ldots, X_m$ are independent and X_i has the same probability distribution as $X_{i-1} + t$, for $i = 2, \ldots, m$, where t is a known positive constant. What is a reasonable modification of the empirical decision rule?

3. In Exercise 2, suppose t is an unknown constant. What is a reasonable modification of the empirical decision rule?

Section 9.1

1. What is the proper modification of the definition of sufficiency given in Sec. 5.13 for conventional statistical problems?

2. Find $\overline{W}(\theta;D;x)$ for each exercise of Sec. 5.6.

Section 9.2

1. Find a minimax test of level of significance 0.2 for the numerical example in the text.

2. Suppose group I contains a single continuous distribution, with joint pdf $f(x;1)$, and group II contains a single continuous distribution, with joint pdf $f(x;2)$. Show that any admissible decision rule has the form: Choose decision 1 for each x for which $f(x;2)/f(x;1) < c$; choose decision 2 for each x for which $f(x;2)/f(x;1) > c$, where c is a constant.

Section 9.3

1. Find a minimax test of level of significance 0.10 for the case where X_1, X_2 are independent, each with an exponential distribution with unknown parameter θ, and $A = 2$, $B = 4$.

2. Find a minimax decision rule of level of significance 0.2 for the case where X_1, X_2, X_3 are independent, each with the probability distribution

Possible values	0	1
Probability	$1-\theta$	θ

and $A = \frac{1}{4}$, $B = \frac{3}{4}$.

Section 9.4

1. Find a minimax test of level of significance 0.05, for the case where X_1, X_2 are independent, each with an exponential distribution with unknown parameter θ and $A = 2$, $B_1 = 1$, $B_2 = 4$.

2. Find a minimax test of level of significance 0.10, for the case where X_1, X_2 are independent, each with a normal distribution with mean zero and unknown standard deviation θ, and $A = 6$, $B_1 = 2$, $B_2 = 9$.

3. Find a minimax test of level of significance 0.15 for the case where X_1, X_2, X_3, X_4 are independent, each with the distribution described in Exercise 2 of Sec. 9.3, and $A = \frac{1}{2}$, $B_1 = \frac{1}{4}$, $B_2 = \frac{3}{4}$.

Section 9.5

1. If $X_1, \ldots, X_m$ are independent, each with the distribution described in Exercise 2 of Sec. 9.3, find the estimate of θ given by the Bayes decision rule relative to the a priori distribution $B(\theta) = \theta$ for $0 < \theta < 1$, where $W(\theta;D) = (D - \theta)^2$.

2. If $X_1, \ldots, X_m$ are independent, each with pdf $f(x;\theta) = (1/\theta)e^{-x/\theta}$ for $x > 0$, $f(x;\theta) = 0$ for $x < 0$, find the estimate of θ given by the Bayes decision rule relative to the cdf $B(\theta) = 1 - e^{-0.01\theta}$ for $\theta > 0$, where $W(\theta;D) = (D - \theta)^2/\theta^2$.

Section 9.6

1. Suppose $X_1, X_2, \ldots$ are independent chance variables, each with a normal distribution with variance equal to 1 and unknown mean θ. The problem is to construct a point estimate of θ, and this estimate can be based on any predetermined number m of the chance variables $X_1, X_2, \ldots$. If m variables are used and the estimate is D, the loss is $5m + 100(D - \theta)^2$. Describe a minimax decision rule for this problem. The rule must specify the value of m and how the decision is made once $X_1, \ldots, X_m$ are observed.

Section 9.7

1. If $X_1, \ldots, X_m$ are all independent, each with pdf $f(x;\theta) = e^{-(x-\theta)}$ for $x > \theta$, $f(x;\theta) = 0$ for $x < \theta$, find the best invariant estimate of θ.

Section 9.8

1. If $X_1, \ldots, X_m$ are all independent, each with pdf $f(x;\theta) = (1/\theta)e^{-(x/\theta)}$ for $x > 0$, $f(x;\theta) = 0$ for $x < 0$, find the best invariant estimate of θ.

Section 9.10

1. If $X_1, \ldots, X_m$ are all independent, each with a uniform distribution between θ_1 and θ_2, where $\theta_1 < \theta_2$, find the maximum-likelihood estimates for θ_1, θ_2.

2. $X_1, \ldots, X_m$ are all independent, each with a normal distribution with known standard deviation σ. $E\{X_i\} = \theta_1 + \theta_2 t_i$, where $t_1, \ldots, t_m$ are given known values, and θ_1, θ_2 are unknown. Find the maximum-likelihood estimates for θ_1, θ_2. Show that these estimates have a joint normal distribution.

3. In Exercise 2, suppose θ_1, θ_2, σ are all unknown. Find the maximum-likelihood estimators for these three parameters.

Section 9.11

1. Suppose $X_1, \ldots, X_m$ are all independent, each with a normal distribution with unknown mean θ_1 and unknown standard deviation θ_2. Group 1 consists of all distributions with $\theta_1 = A$, where A is a given value. Find the likelihood-ratio rule for testing this hypothesis. Show that a table of the t distribution can be used to help specify the rule.

2. Suppose $X_1, \ldots, X_5$ are all independent, each with a uniform distribution between θ_1 and θ_2, where $\theta_1 < \theta_2$. Group 1 consists of all distributions with $\theta_2 - \theta_1 < 2$. Find the likelihood-ratio rule of level of significance 0.1 for testing this hypothesis.

APPENDIX

TABLE 1

This table gives the value of $(1/\sqrt{2\pi})\int_{-\infty}^{z} e^{-(1/2)y^2}\,dy$ for the values of z listed. The value of z for any cell is given by adding the number at the extreme left of the row in which the cell appears to the number at the extreme top of the column in which the cell appears. The value of the integral for a negative z can be found by subtracting its value for $|z|$ from 1.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0 9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0 9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0 9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0 9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

TABLE 2

If W has a chi-square distribution with n degrees of freedom, this table gives the value of w for which $P(W \leqslant w) = A$, for the values of n listed in the column at the left of the table and the values of A listed in the row at the top of the table.

Value of n	Value of A												
	0.01	0.02	0.05	0.10	0.20	0.30	0.50	0.70	0.80	0.90	0.95	0.98	0.99
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

SOURCE: Reprinted from R. A. Fisher, "Statistical Methods for Research Workers," Table III, Oliver & Boyd Ltd., Edinburgh, by permission of the author and publishers.

TABLE 3

If *T* has a *t* distribution with *n* degrees of freedom, this table gives the value of *t* for which $P(-t \leqslant T \leqslant t) = A$, for the values of *n* listed in the column at the left of the table and the values of *A* listed in the row at the top of the table.

Value of *n*	Value of *A*											
	0.99	0.98	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
1	63.657	31.821	12.706	6.314	3.078	1.963	1.376	1.000	0.727	0.510	0.325	0.158
2	9.925	6.965	4 303	2.920	1.886	1.386	1.061	0.816	0.617	0.445	0.289	0.142
3	5.841	4.541	3.182	2.353	1.638	1.250	0.978	0.765	0.584	0 424	0 277	0.137
4	4.604	3.747	2.776	2.132	1.533	1.190	0.941	0.741	0.569	0.414	0.271	0.134
5	4.032	3.365	2 571	2.015	1.476	1.156	0.920	0.727	0.559	0.408	0.267	0.132
6	3.707	3.143	2.447	1.943	1.440	1.134	0.906	0.718	0.553	0.404	0.265	0.131
7	3.499	2.998	2.365	1.895	1.415	1.119	0.896	0.711	0.549	0.402	0.263	0.130
8	3.355	2.896	2.306	1.860	1.397	1.108	0.889	0 706	0.546	0.399	0.262	0.130
9	3.250	2.821	2.262	1.833	1.383	1.100	0.883	0.703	0.543	0.398	0.261	0.129
10	3.169	2.764	2.228	1.812	1.372	1.093	0.879	0.700	0.542	0.397	0.260	0.129
11	3.106	2.718	2 201	1.796	1.363	1.088	0.876	0.697	0.540	0.396	0.260	0.129
12	3.055	2.681	2.179	1.782	1.356	1 083	0 873	0 695	0.539	0.395	0.259	0.128
13	3.012	2.650	2.160	1.771	1.350	1.079	0.870	0.694	0 538	0.394	0.259	0.128
14	2.977	2.624	2.145	1.761	1.345	1.076	0.868	0.692	0.537	0.393	0.258	0.128
15	2.947	2.602	2.131	1.753	1.341	1.074	0.866	0.691	0.536	0.393	0.258	0.128
16	2 921	2 583	2.120	1.746	1.337	1.071	0.865	0.690	0.535	0.392	0.258	0.128
17	2.898	2.567	2.110	1.740	1.333	1.069	0.863	0.689	0.534	0.392	0.257	0.128
18	2.878	2.552	2.101	1.734	1.330	1.067	0.862	0.688	0.534	0.392	0.257	0.127
19	2.861	2.539	2.093	1.729	1.328	1.066	0.861	0.688	0.533	0.391	0.257	0.127
20	2.845	2.528	2.086	1.725	1.325	1.064	0.860	0.687	0.533	0.391	0.257	0.127
21	2.831	2.518	2.080	1.721	1.323	1.063	0.859	0.686	0.532	0.391	0.257	0.127
22	2.819	2 508	2.074	1.717	1.321	1.061	0.858	0.686	0 532	0.390	0.256	0.127
23	2.807	2.500	2.069	1.714	1.319	1.060	0.858	0.685	0.532	0.390	0.256	0.127
24	2.797	2.492	2.064	1.711	1.318	1.059	0.857	0 685	0.531	0.390	0.256	0.127
25	2.787	2.485	2.060	1.708	1.316	1.058	0.856	0 684	0.531	0.390	0.256	0.127
26	2.779	2.479	2.056	1.706	1.315	1.058	0.856	0.684	0.531	0.390	0.256	0.127
27	2.771	2.473	2.052	1.703	1.314	1.057	0.855	0.684	0.531	0.389	0.256	0.127
28	2.763	2.467	2.048	1.701	1.313	1.056	0.855	0 683	0.530	0.389	0.256	0.127
29	2.756	2.462	2.045	1.699	1.311	1.055	0.854	0 683	0.530	0.389	0.256	0.127
30	2.750	2.457	2.042	1.697	1.310	1.055	0.854	0.683	0.530	0.389	0.256	0.127
∞	2.576	2.326	1.960	1.645	1.282	1.036	0.842	0 674	0.524	0.385	0.253	0.126

SOURCE: Reprinted from R. A. Fisher, "Statistical Methods for Research Workers," Table IV, Oliver & Boyd Ltd., Edinburgh, by permission of the author and publishers.

TABLE 4

If Z has an F distribution with r degrees of freedom in the numerator and s degrees of freedom in the denominator, this table gives the value of z for which $P(Z \geq z) = 0.05$ (in regular type) and the value of z for which $P(Z \geq z) = 0.01$ (in boldface type), for the values of r listed along the top of the table and the values of s listed along the side of the table.

Value of s	Value of r																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**	245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,258**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**
3	10.13 **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**	2.42 **3.59**	2.38 **3.51**	2.34 **3.42**	2.32 **3.37**	2.28 **3.30**	2.26 **3.27**	2.24 **3.21**	2.22 **3.18**	2.21 **3.16**
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**	2.48 **3.70**	2.44 **3.62**	2.39 **3.51**	2.35 **3.43**	2.31 **3.34**	2.27 **3.26**	2.24 **3.21**	2.21 **3.14**	2.19 **3.11**	2.16 **3.06**	2.14 **3.02**	2.13 **3.00**
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.70 **4.14**	2.64 **4.00**	2.59 **3.89**	2.55 **3.80**	2.51 **3.73**	2.48 **3.67**	2.43 **3.56**	2.39 **3.48**	2.33 **3.36**	2.29 **3.29**	2.25 **3.20**	2.21 **3.12**	2.18 **3.07**	2.15 **3.00**	2.12 **2.97**	2.10 **2.92**	2.08 **2.89**	2.07 **2.87**
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.66 **4.03**	2.59 **3.89**	2.54 **3.78**	2.49 **3.69**	2.45 **3.61**	2.42 **3.55**	2.37 **3.45**	2.33 **3.37**	2.28 **3.25**	2.24 **3.18**	2.20 **3.10**	2.16 **3.01**	2.13 **2.96**	2.09 **2.89**	2.07 **2.86**	2.04 **2.80**	2.02 **2.77**	2.01 **2.75**

17	4.45 8.40	3.59 6.11	3.20 5.18	2.96 4.67	2.81 4.34	2.70 4.10	2.62 3.93	2.55 3.79	2.50 3.68	2.45 3.59	2.41 3.52	2.38 3.45	2.33 3.35	2.29 3.27	2.23 3.16	2.19 3.08	2.15 3.00	2.11 2.92	2.08 2.86	2.04 2.79	2.02 2.76	1.99 2.70	1.97 2.67	1.96 2.65
18	4.41 8.28	3.55 6.01	3.16 5.09	2.93 4.58	2.77 4.25	2.66 4.01	2.58 3.85	2.51 3.71	2.46 3.60	2.41 3.51	2.37 3.44	2.34 3.37	2.29 3.27	2.25 3.19	2.19 3.07	2.15 3.00	2.11 2.91	2.07 2.83	2.04 2.78	2.00 2.71	1.98 2.68	1.95 2.62	1.93 2.59	1.92 2.57
19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50	2.74 4.17	2.63 3.94	2.55 3.77	2.48 3.63	2.43 3.52	2.38 3.43	2.34 3.36	2.31 3.30	2.26 3.19	2.21 3.12	2.15 3.00	2.11 2.92	2.07 2.84	2.02 2.76	2.00 2.70	1.96 2.63	1.94 2.60	1.91 2.54	1.90 2.51	1.88 2.49
20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43	2.71 4.10	2.60 3.87	2.52 3.71	2.45 3.56	2.40 3.45	2.35 3.37	2.31 3.30	2.28 3.23	2.23 3.13	2.18 3.05	2.12 2.94	2.08 2.86	2.04 2.77	1.99 2.69	1.96 2.63	1.92 2.56	1.90 2.53	1.87 2.47	1.85 2.44	1.84 2.42
21	4.32 8.02	3.47 5.78	3.07 4.87	2.84 4.37	2.68 4.04	2.57 3.81	2.49 3.65	2.42 3.51	2.37 3.40	2.32 3.31	2.28 3.24	2.25 3.17	2.20 3.07	2.15 2.99	2.09 2.88	2.05 2.80	2.00 2.72	1.96 2.63	1.93 2.58	1.89 2.51	1.87 2.47	1.84 2.42	1.82 2.38	1.81 2.36
22	4.30 7.94	3.44 5.72	3.05 4.82	2.82 4.31	2.66 3.99	2.55 3.76	2.47 3.59	2.40 3.45	2.35 3.35	2.30 3.26	2.26 3.18	2.23 3.12	2.18 3.02	2.13 2.94	2.07 2.83	2.03 2.75	1.98 2.67	1.93 2.58	1.91 2.53	1.87 2.46	1.84 2.42	1.81 2.37	1.80 2.33	1.78 2.31
23	4.28 7.88	3.42 5.66	3.03 4.76	2.80 4.26	2.64 3.94	2.53 3.71	2.45 3.54	2.38 3.41	2.32 3.30	2.28 3.21	2.24 3.14	2.20 3.07	2.14 2.97	2.10 2.89	2.04 2.78	2.00 2.70	1.96 2.62	1.91 2.53	1.88 2.48	1.84 2.41	1.82 2.37	1.79 2.32	1.77 2.28	1.76 2.26
24	4.26 7.82	3.40 5.61	3.01 4.72	2.78 4.22	2.62 3.90	2.51 3.67	2.43 3.50	2.36 3.36	2.30 3.25	2.26 3.17	2.22 3.09	2.18 3.03	2.13 2.93	2.09 2.85	2.02 2.74	1.98 2.66	1.94 2.58	1.89 2.49	1.86 2.44	1.82 2.36	1.80 2.33	1.76 2.27	1.74 2.23	1.73 2.21
25	4.24 7.77	3.38 5.57	2.99 4.68	2.76 4.18	2.60 3.86	2.49 3.63	2.41 3.46	2.34 3.32	2.28 3.21	2.24 3.13	2.20 3.05	2.16 2.99	2.11 2.89	2.06 2.81	2.00 2.70	1.96 2.62	1.92 2.54	1.87 2.45	1.84 2.40	1.80 2.32	1.77 2.29	1.74 2.23	1.72 2.19	1.71 2.17
26	4.22 7.72	3.37 5.53	2.89 4.64	2.74 4.14	2.59 3.82	2.47 3.59	2.39 3.42	2.32 3.29	2.27 3.17	2.22 3.09	2.18 3.02	2.15 2.96	2.10 2.86	2.05 2.77	1.99 2.66	1.95 2.58	1.90 2.50	1.85 2.41	1.82 2.36	1.78 2.28	1.76 2.25	1.72 2.19	1.70 2.15	1.69 2.13
27	4.21 7.68	3.35 5.49	2.96 4.60	2.73 4.11	2.57 3.79	2.46 3.56	2.37 3.39	2.30 3.26	2.25 3.14	2.20 3.06	2.16 2.98	2.13 2.93	2.08 2.83	2.03 2.74	1.97 2.63	1.93 2.55	1.88 2.47	1.84 2.38	1.80 2.33	1.76 2.25	1.74 2.21	1.71 2.16	1.68 2.12	1.67 2.10
28	4.20 7.64	3.34 5.45	2.95 4.57	2.71 4.07	2.56 3.76	2.44 3.53	2.36 3.36	2.29 3.23	2.24 3.11	2.19 3.03	2.15 2.95	2.12 2.90	2.06 2.80	2.02 2.71	1.96 2.60	1.91 2.52	1.87 2.44	1.81 2.35	1.78 2.30	1.75 2.22	1.72 2.18	1.69 2.13	1.67 2.09	1.65 2.06
29	4.18 7.60	3.33 5.42	2.93 4.54	2.70 4.04	2.54 3.73	2.43 3.50	2.35 3.33	2.28 3.20	2.22 3.08	2.18 3.00	2.14 2.92	2.10 2.87	2.05 2.77	2.00 2.68	1.94 2.57	1.90 2.49	1.85 2.41	1.80 2.32	1.77 2.27	1.73 2.19	1.71 2.15	1.68 2.10	1.65 2.06	1.64 2.03
30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70	2.42 3.47	2.34 3.30	2.27 3.17	2.21 3.06	2.16 2.98	2.12 2.90	2.09 2.84	2.04 2.74	1.99 2.66	1.93 2.55	1.89 2.47	1.84 2.38	1.79 2.29	1.76 2.24	1.72 2.16	1.69 2.13	1.66 2.07	1.64 2.03	1.62 2.01
32	4.15 7.50	3.30 5.34	2.90 4.46	2.67 3.97	2.51 3.66	2.40 3.42	2.32 3.25	2.25 3.12	2.19 3.01	2.14 2.94	2.10 2.86	2.07 2.80	2.02 2.70	1.97 2.62	1.91 2.51	1.86 2.42	1.82 2.34	1.76 2.25	1.74 2.20	1.69 2.12	1.67 2.08	1.64 2.02	1.61 1.98	1.59 1.96
34	4.13 7.44	3.28 5.29	2.88 4.42	2.65 3.93	2.49 3.61	2.38 3.38	2.30 3.21	2.23 3.08	2.17 2.97	2.12 2.89	2.08 2.82	2.05 2.76	2.00 2.66	1.95 2.58	1.89 2.47	1.84 2.38	1.80 2.30	1.74 2.21	1.71 2.15	1.67 2.08	1.64 2.04	1.61 1.98	1.59 1.94	1.57 1.91
36	4.11 7.39	3.26 5.25	2.86 4.38	2.63 3.89	2.48 3.58	2.36 3.35	2.28 3.18	2.21 3.04	2.15 2.94	2.10 2.86	2.06 2.78	2.03 2.72	1.98 2.62	1.93 2.54	1.87 2.43	1.82 2.35	1.78 2.26	1.72 2.17	1.69 2.12	1.65 2.04	1.62 2.00	1.59 1.94	1.56 1.90	1.55 1.87
38	4.10 7.35	3.25 5.21	2.85 4.34	2.62 3.86	2.46 3.54	2.35 3.32	2.26 3.15	2.19 3.02	2.14 2.91	2.09 2.82	2.05 2.75	2.02 2.69	1.96 2.59	1.92 2.51	1.85 2.40	1.80 2.32	1.76 2.22	1.71 2.14	1.67 2.08	1.63 2.00	1.60 1.97	1.57 1.90	1.54 1.86	1.53 1.84
40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51	2.34 3.29	2.25 3.12	2.18 2.99	2.12 2.88	2.07 2.80	2.04 2.73	2.00 2.66	1.95 2.56	1.90 2.49	1.84 2.37	1.79 2.29	1.74 2.20	1.69 2.11	1.66 2.05	1.61 1.97	1.59 1.94	1.55 1.88	1.53 1.84	1.51 1.81
42	4.07 7.27	3.22 5.15	2.83 4.29	2.59 3.80	2.44 3.49	2.32 3.26	2.24 3.10	2.17 2.96	2.11 2.86	2.06 2.77	2.02 2.70	1.99 2.64	1.94 2.54	1.89 2.46	1.82 2.35	1.78 2.26	1.73 2.17	1.68 2.08	1.64 2.02	1.60 1.94	1.57 1.91	1.54 1.85	1.51 1.80	1.49 1.78

TABLE 4 (*concluded*)

Value of *s*	Value of *r*																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
44	4.06 7.24	3.21 5.12	2.82 4.26	2.58 3.78	2.43 3.46	2.31 3.24	2.23 3.07	2.16 2.94	2.10 2.84	2.05 2.75	2.01 2.68	1.98 2.62	1.92 2.52	1.88 2.44	1.81 2.32	1.76 2.24	1.72 2.15	1.66 2.06	1.63 2.00	1.58 1.92	1.56 1.88	1.52 1.82	1.50 1.78	1.48 1.75
46	4.05 7.21	3.20 5.10	2.81 4.24	2.57 3.76	2.42 3.44	2.30 3.22	2.22 3.05	2.14 2.92	2.09 2.82	2.04 2.73	2.00 2.66	1.97 2.60	1.91 2.50	1.87 2.42	1.80 2.30	1.75 2.22	1.71 2.13	1.65 2.04	1.62 1.98	1.57 1.90	1.54 1.86	1.51 1.80	1.48 1.76	1.46 1.72
48	4.04 7.19	3.19 5.08	2.80 4.22	2.56 3.74	2.41 3.42	2.30 3.20	2.21 3.04	2.14 2.90	2.08 2.80	2.03 2.71	1.99 2.64	1.96 2.58	1.90 2.48	1.86 2.40	1.79 2.28	1.74 2.20	1.70 2.11	1.64 2.02	1.61 1.96	1.56 1.88	1.53 1.84	1.50 1.78	1.47 1.73	1.45 1.70
50	4.03 7.17	3.18 5.06	2.79 4.20	2.56 3.72	2.40 3.41	2.29 3.18	2.20 3.02	2.13 2.88	2.07 2.78	2.02 2.70	1.98 2.62	1.95 2.56	1.90 2.46	1.85 2.39	1.78 2.26	1.74 2.18	1.69 2.10	1.63 2.00	1.60 1.94	1.55 1.86	1.52 1.82	1.48 1.76	1.46 1.71	1.44 1.68
55	4.02 7.12	3.17 5.01	2.78 4.16	2.54 3.68	2.38 3.37	2.27 3.15	2.18 2.98	2.11 2.85	2.05 2.75	2.00 2.66	1.97 2.59	1.93 2.53	1.88 2.43	1.83 2.35	1.76 2.23	1.72 2.15	1.67 2.06	1.61 1.96	1.58 1.90	1.52 1.82	1.50 1.78	1.46 1.71	1.43 1.66	1.41 1.64
60	4.00 7.08	3.15 4.98	2.76 4.13	2.52 3.65	2.37 3.34	2.25 3.12	2.17 2.95	2.10 2.82	2.04 2.72	1.99 2.63	1.95 2.56	1.92 2.50	1.86 2.40	1.81 2.32	1.75 2.20	1.70 2.12	1.65 2.03	1.59 1.93	1.56 1.87	1.50 1.79	1.48 1.74	1.44 1.68	1.41 1.63	1.39 1.60
65	3.99 7.04	3.14 4.95	2.75 4.10	2.51 3.62	2.36 3.31	2.24 3.09	2.15 2.93	2.08 2.79	2.02 2.70	1.98 2.61	1.94 2.54	1.90 2.47	1.85 2.37	1.80 2.30	1.73 2.18	1.68 2.09	1.63 2.00	1.57 1.90	1.54 1.84	1.49 1.76	1.46 1.71	1.42 1.64	1.39 1.60	1.37 1.56
70	3.98 7.01	3.13 4.92	2.74 4.08	2.50 3.60	2.35 3.29	2.32 3.07	2.14 2.91	2.07 2.77	2.01 2.67	1.97 2.59	1.93 2.51	1.89 2.45	1.84 2.35	1.79 2.28	1.72 2.15	1.67 2.07	1.62 1.98	1.56 1.88	1.53 1.82	1.47 1.74	1.45 1.69	1.40 1.62	1.37 1.56	1.35 1.53
80	3.96 6.96	3.11 4.88	2.72 4.04	2.48 3.56	2.33 3.25	2.21 3.04	2.12 2.87	2.05 2.74	1.99 2.64	1.95 2.55	1.91 2.48	1.88 2.41	1.82 2.32	1.77 2.24	1.70 2.11	1.65 2.03	1.60 1.94	1.54 1.84	1.51 1.78	1.45 1.70	1.42 1.65	1.38 1.57	1.35 1.52	1.32 1.49
100	3.94 6.90	3.09 4.82	2.70 3.98	2.46 3.51	2.30 3.20	2.19 2.99	2.10 2.82	2.03 2.69	1.97 2.59	1.92 2.51	1.88 2.43	1.85 2.36	1.79 2.26	1.75 2.19	1.68 2.06	1.63 1.98	1.57 1.89	1.51 1.79	1.48 1.73	1.42 1.64	1.39 1.59	1.34 1.51	1.30 1.46	1.28 1.43
125	3.92 6.84	3.07 4.78	2.68 3.94	2.44 3.47	2.29 3.17	2.17 2.95	2.08 2.79	2.01 2.65	1.95 2.56	1.90 2.47	1.86 2.40	1.83 2.33	1.77 2.23	1.72 2.15	1.65 2.03	1.60 1.94	1.55 1.85	1.49 1.75	1.45 1.68	1.39 1.59	1.36 1.54	1.31 1.46	1.27 1.40	1.25 1.37
150	3.91 6.81	3.06 4.75	2.67 3.91	2.43 3.44	2.27 3.13	2.16 2.92	2.07 2.76	2.00 2.62	1.94 2.53	1.89 2.44	1.85 2.37	1.82 2.30	1.76 2.20	1.71 2.12	1.64 2.00	1.59 1.91	1.54 1.83	1.47 1.72	1.44 1.66	1.37 1.56	1.34 1.51	1.29 1.43	1.25 1.37	1.22 1.33
200	3.89 6.76	3.04 4.71	2.65 3.88	2.41 3.41	2.26 3.11	2.14 2.90	2.05 2.73	1.98 2.60	1.92 2.50	1.87 2.41	1.83 2.34	1.80 2.28	1.74 1.17	1.69 2.09	1.62 1.97	1.57 1.88	1.52 1.79	1.45 1.69	1.42 1.62	1.35 1.53	1.32 1.48	1.26 1.39	1.22 1.33	1.19 1.28
400	3.86 6.70	3.02 4.66	2.62 3.83	2.39 3.36	2.23 3.06	2.12 2.85	2.03 2.69	1.96 2.55	1.90 2.46	1.85 2.37	1.81 2.29	1.78 2.23	1.72 2.12	1.67 2.04	1.60 1.92	1.54 1.84	1.49 1.74	1.42 1.64	1.38 1.57	1.32 1.47	1.28 1.42	1.22 1.32	1.16 1.24	1.13 1.19
1,000	3.85 6.66	3.00 4.62	2.61 3.80	2.38 3.34	2.22 3.04	2.10 2.82	2.02 2.66	1.95 2.53	1.89 2.43	1.84 2.34	1.80 2.26	1.76 2.20	1.70 2.09	1.65 2.01	1.58 1.89	1.53 1.81	1.47 1.71	1.41 1.61	1.36 1.54	1.30 1.44	1.26 1.38	1.19 1.28	1.13 1.19	1.08 1.11
∞	3.84 6.64	2.99 4.60	2.60 3.78	2.37 3.32	2.21 3.02	2.09 2.80	2.01 2.64	1.94 2.51	1.88 2.41	1.83 2.32	1.79 2.24	1.75 2.18	1.69 2.07	1.64 1.99	1.57 1.87	1.52 1.79	1.46 1.69	1.40 1.59	1.35 1.52	1.28 1.41	1.24 1.36	1.17 1.25	1.11 1.15	1.00 1.00

SOURCE: Reproduced by permission from George W. Snedecor, "Statistical Methods," 5th ed., 1956, copyright, Iowa State University Press, Ames, Iowa.

INDEX